CHILD'S PLAY: ACTIVITY RECOGNITION FOR MONITORING CHILDREN'S DEVELOPMENTAL PROGRESS WITH AUGMENTED TOYS

A Thesis
Presented to
The Academic Faculty

by

Tracy L. Westeyn

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
College of Computing,
School of Interactive Computing

Georgia Institute of Technology
August 2010

Copyright © 2010 by Tracy L. Westeyn

CHILD'S PLAY: ACTIVITY RECOGNITION FOR MONITORING CHILDREN'S DEVELOPMENTAL PROGRESS WITH AUGMENTED TOYS

Approved by:

Professor Thad E. Starner, Ph.D.,
Advisor
College of Computing,
School of Interactive Computing
Georgia Institute of Technology

Professor Paul Lukowicz, Ph.D
Department of Mathematics and
Informatics
University of Passu, Germany

Professor Gregory D. Abowd, Ph.D
Co–Advisor
College of Computing,
School of Interactive Computing
Georgia Institute of Technology

Professor Melody Moore Jackson, Ph.D
College of Computing,
School of Interactive Computing
Georgia Institute of Technology

Professor James M. Rehg, Ph.D
College of Computing,
School of Interactive Computing
Georgia Institute of Technology

Rosa Arriaga, Ph.D.
College of Computing,
School of Interactive Computing
Georgia Institute of Technology

Date Approved: May 10, 2010

For my family...

ACKNOWLEDGEMENTS

I would like to thank my advisors, Thad Starner and Gregory Abowd for the continued support of this research. I would also like to thank my thesis committee for the advice and time. I would like to thank Dr. Grace Baranek, Dr. Lauren Adamson, and Dr. Agata Rozga for the early guidance with this research. In addition, I would also like to thank my colleagues in the Contextual Computing Group, the Ubiquitous Computing Group, and in the Interactive Media and Technology Center. I would especially like to thank Jeremy Johnson, Peter Presti, David Minnen, Allison Elliot Tew, Julie Kientz, Helene Brashear, Kim Weaver, Daniel Gifford, Jiasheng He, David Quigley, Scott Gilliland, and Valerie Summet.

This work has been supported in part by the Achievement Awards for College Scientists Atlanta Chapter, Children's Health Care of Atlanta, the Health Systems Institute, Autism Speaks, and the Graphics Visualization and Usability Center. Development of the wireless sensing technology was funded in part by the National Science Foundation and the National Institute on Disability and Rehabilitation Research. This material is based upon work supported by the National Science Foundation (NSF) under Grant No. 0812281.[1] The Rehabilitation Engineering Research Center for Wireless Technologies is sponsored by the National Institute on Disability and Rehabilitation Research (NIDRR) of the U.S. Department of Education under grant number H133E060061. [2]

[1]Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of NSF.

[2]The opinions contained in this material are those of the author and do not necessarily reflect those of the U.S. Department of Education or NIDRR.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The way in which infants play with objects can be indicative of their developmental progress and may serve as an early indicator for developmental delays. However, the observation of children interacting with toys for the purpose of quantitative analysis can be a difficult task. To better quantify how play may serve as an early indicator, researchers have conducted retrospective studies examining the differences in object play behaviors among infants. However, such studies require that researchers repeatedly inspect videos of play often at speeds much slower than real-time to indicate points of interest. The research presented in this dissertation examines whether a combination of sensors embedded within toys and automatic pattern recognition of object play behaviors can help expedite this process.

For my dissertation, I developed the *Child'sPlay* system which uses augmented toys and statistical models to automatically provide quantitative measures of object play interactions, as well as, provide the *PlayView* interface to view annotated play data for later analysis. In this dissertation, I examine the hypothesis that sensors embedded in objects can provide sufficient data for automatic recognition of certain exploratory, relational, and functional object play behaviors in semi-naturalistic environments and that a continuum of recognition accuracy exists which allows automatic indexing to be useful for retrospective review.

I designed several augmented toys and used them to collect object play data from more than fifty play sessions. I conducted pattern recognition experiments over this data to produce statistical models that automatically classify children's object play behaviors. In addition, I conducted a user study with twenty participants to determine if annotations automatically generated from these models help improve performance in retrospective review tasks. My results indicate that these statistical models increase user performance and decrease perceived effort when combined with the *PlayView* interface during retrospective review. The presence of high quality annotations are preferred by users and promotes an increase in the effective retrieval rates of object play behaviors.

CHAPTER I

INTRODUCTION AND MOTIVATION

The observation of infants' and toddlers' developmental progress for the purpose of quantitative analysis is an important and yet difficult task. A developmental delay is diagnosed when a child fails to exhibit behavioral milestones typical for their age group. The prevalence of developmental delay in the United States for young children is approximately 10 percent [19]. As such, the early identification of these children is an important public health goal. However, the wide variation of typical development among children can make establishing the presence of a developmental delay a difficult task. Subtle, early abnormalities can often be overlooked as normal developmental variation [9]. This issue is further exacerbated by the large number of markers used to track developmental progress. The routine monitoring of a child's progress is crucial to the identification of delays and is considered a vital component of pediatric care [63]. Recent formative research has explored the monitoring practices and record-keeping needs for families of young children. The results suggest that parents may benefit from increased support with the manual tracking of their child's developmental progress and that mobile ubiquitous computer technology may help in this domain [38]. These devices may benefit from further development of technology and algorithms that can automate the identification and recording of developmentally relevant activities. Such automated technology has yet to be explored.

The Centers for Disease Control and Prevention currently lists over 200 milestones to track over the first five years of a child's life [13]. These developmental skills can range from banging a toy on a table to displaying socially appropriate expressions to siblings. With developmental milestones spanning the four main areas of cognitive, physical, linguistic, and psychosocial skills, a large variety of mobile sensors can be employed to collect data. However, automatically recognizing all developmental milestones across all four areas is a task beyond the scope of this dissertation. Specialists often use a subset of these milestones

in screening diagnostics, and recent research in Psychology suggests that the observation of object play interactions may help identify early indicators of certain developmental delays [2, 7]. A subset of play activities, similar to those studied in clinical research, will be the focus of the automatic recognition aspects of this work.

The ways in which infants and toddlers play with objects can be indicative of their developmental progress. Depending on their age, a child's object play activities can display simple physical milestones such as placing objects in their mouth to sophisticated cognitive tasks such as symbolically using a banana as a telephone receiver. Psychologists have created a coding scheme which quantizes the levels of sophistication displayed by infants while engaged in object play (see Section 2.1.1 for more detail). Using retrospective analysis of home videos for children diagnosed with autism spectrum disorder, Baranek *et al.* used this scheme to identify the highest level of sophistication reached per child and found that the level and duration of play at each level differed between typically developing children and children with autism [7]. Object play behaviors offer a viable subset of developmental milestones to explore for the purposes of automatic recognition.

This research will produce technology that can automatically generate quantitative data from observations of children engaged in object play, similar to that produced by the coding scheme of Baranek *et al.* These measures include the frequency with which an object is played, the time spent attending between different objects, and the highest level of play sophistication reached by a child. The technology presented in this dissertation focuses on recognition within the first six levels of sophistication described by this scheme which include play behaviors from exploratory, relational, and functional play. These categories include toy manipulations such as grasping, shaking, rolling a ball, pushing, pulling apart interlocking toys, uncovering lids, pouring, stacking blocks, and early imaginary actions.

Embedding wireless sensors in toys may allow the infant object play to control computational objects automatically within the environment to help collect relevant data and promote the future study of these interactions. For example, with this technology a video capable ultra mobile device, such as KidCam [37], could automatically save video footage as a toddler assembles blocks, babbles while removing lids from containers, or achieves some

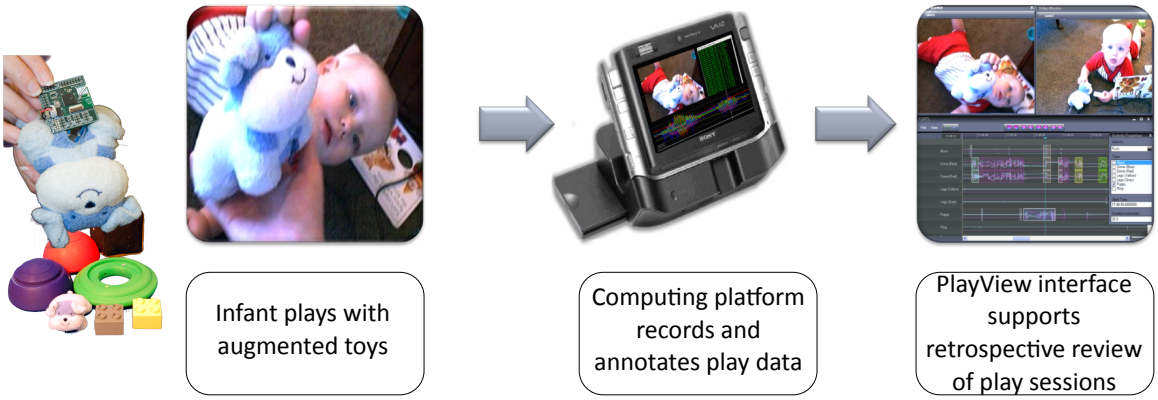


Figure 1: Components of the *Child'sPlay* system

other developmental milestone. This collection of relevant video data (as opposed to a large continuous stream of video footage) may allow for both a more rapid and more frequent analysis of developmental play activities by researchers. As a proof of concept, I present technology focused on *in situ* controlled studies. However, in the future these technologies could be extended to generalize to home use.

The automatic detection of developmental indicators is an interesting pattern recognition problem. This dissertation presents an applied algorithm for automatically recognizing object play primitives and their composition to identify higher levels of play sophistication along with the design and implementation of wireless sensor enabled toys. In addition, this dissertation also explores the challenges associated with collecting play data from young children and the viability of using play data gathered from adults to generate statistical models that can then be applied to recognize play behaviors among young children. As part of this exploration, I collect a data set of adult object play and compare the ability of various algorithms to compose primitive play behaviors into higher levels of play sophistication. In particular, I explore boosting, hidden Markov models (HMMs), and support vector machines as well as naïve methods. Knowing that, regardless of the algorithm used, automatic recognition of play is not perfect, this dissertation also investigates the impact that recognition errors have on the ability of a user to interact with an intelligent interface designed for retrospective review of object play behaviors. This investigation also provides a target in terms of recognition accuracy for future systems.

1.1 Purpose of Research

The goal of this research is to develop components of technological tools that could one day help increase the understanding of children's development through the automatic capture, access, and retrospective review of toddler-object play in semi-structured environments. This research includes the exploration of a wireless sensor solution consisting of sensors embedded in toys, a statistical pattern recognition component that can automatically characterize certain types of object play behavior performed by young children, and an interface which supports retrospective review of play data (see Figure 1). In particular, this technology will automatically identify rudimentary object play behaviors and characterize the types of play observed via the *PlayView* interface.

1.2 Thesis Statement

I hypothesize that sensors embedded in objects can provide sufficient data for automatic recognition of certain exploratory, relational, and functional object play behaviors in semi-naturalistic environments and that a continuum of recognition accuracy exists which allows automatic indexing to be useful for retrospective review.

.

1.3 Research Questions

This thesis will explore the following research questions:

1. Can wireless sensors embedded in toys be used to collect sufficient data in semi–naturalistic settings to automatically identify specific exploratory, relational, and functional play behaviors?

2. Can statistical models of object play developed with adult play data be generalized to allow recognition of a child's object play with sufficient enough quality to support later review by other individuals?

3. What level of recognition quality do users find acceptable when using systems to support retrospective analysis of object play behaviors?

1.4 Research Contributions

The following contributions are made during the process of exploring the thesis statement in Section 1.2:

1. The first contribution is an exploration of the use of wireless sensors embedded in age–appropriate toys to support the transparent capture of quantitative toddler–object interaction data in naturalistic settings. As part of this contribution, I also designed and developed toys that allow for sensors to be easily shared and exchanged for those with other sensing modalities.

2. The second contribution is the application and exploration of statistical machine learning techniques to learn and automatically recognize operational definitions of exploratory, relational, and functional play. As part of this contribution, I investigate how well these techniques generalize across different participants and toys as well as characterize the strengths and weaknesses of each technique for identifying object play.

3. The third contribution is the production of a consistent method for coding developmental data by using the recognized play behaviors to automatically index associated video footage and object play data for later analysis by other individuals.

4. The fourth contribution is an exploration of the impact that recognition quality has on the user–experience when annotating object play data via an interface designed for retrospective review.

5. The fifth contribution is the production of multivariate, multiple modality data sets of object play behavior collected from adults, young children, and toddlers to help encourage further pattern recognition research in this area.

1.5 Thesis Overview

Chapter 2 presents background, which supports the need for the results of this research by covering five distinct areas of work related to this thesis: Psychology research on object play; capture–and–access systems for children with disabilities; existing sensor packages along with pattern recognition techniques for activities of daily living; evaluation metrics for continuous activity recognition; and augmented toy systems. In Chapter 3, I present my previous work with on–body sensing systems designed to collect data for later review. This chapter also describes the pattern recognition methods used in my previous work as well as highlighting evaluation methods for continuous activity recognition. Chapter 4 describes the design considerations for both the sensing technology and the toys used to collect object play data. I also discuss the initial play tests of the prototype toys that informed the design of the final toy set. Chapter 5 details a pilot study conducted late in 2007 and provides initial recognition results of adults playing with a mixture of augmented and regular off–the–shelf toys. Based on the pilot study detailed in Chapter 5, I present motivation for restricting my pattern recognition experiments towards play data using only augmented toys. Chapter 6 describes both the adult and child data sets collected from Fall 2008 until Fall 2009, providing details on pattern recognition experiments on both adult and child data sets. Knowing that recognition rates will not be perfect, in Chapter 7, I detail the results of a study that assesses the impact that recognition errors have on the ability of a user to identify object play behaviors using a data visualization tool. Chapter 8 and Chapter 9 present a discussion of future directions for this work and my conclusions.

CHAPTER II

BACKGROUND AND RELATED WORK

In this chapter, I discuss some of the background and work related to key areas of this thesis including: exploring developmental progress via infant object play, capture–and–access systems for children with disabilities, pattern recognition systems for activities of daily living using wireless sensors, evaluation of continuous activity recognition systems, and augmented toy systems as tangible interfaces.

2.1 Studying Developmental Progress via Object Play

Infant–object interaction has been a focus of study since the early 1920s [65]. There are many different types of object play that a researcher may wish to explore when attempting to identify early indicators of developmental delays. Some current areas being explored clinically are: social aspects of object play, such as sharing attention between objects and playmates; physical manipulation aspects of object play, such as stacking objects; and attentional aspects of object play, such as object preferences [2, 7, 65, 35]. My thesis directly builds on work in physical manipulation object play and shares similar procedural elements with work conducted in shared attention studies involving object play.

2.1.1 Levels of Sophistication in Object Play

There has been much research in Psychology investigating identification of early indicators for a wide variety of developmental delays [19]. Of particular interest is a video retrospective analysis performed by Baranek *et al.* correlating infants' interactions with inanimate objects (called object play) with the child's current diagnoses as either typical, having an autism spectrum disorder, or otherwise developmentally delayed [7]. This study provided the first exploration of object play for infants between the ages of 9–12 months.

To perform this analysis, Baranek *et al.* elicited home movies of 32 infants from families whose children were now old enough to receive a diagnosis indicating if they are typical

or atypical with respect to development (and, in particular, a diagnosis of autism). The content of these videos was then coded for high level scene information such as the scene location, the number of people in the scene, and the presence of infants. This gross content analysis was then used to composite clips of the infants from a variety of scenes into a 10 minute sampling of the video for each infant. These 10 minute samplings were then coded for object play to determine the play ability demonstrated by each child.

Prior to the study by Baranek *et al.*, there was no universally accepted scale to rate object play. In order to provide a measure of play ability, Baranek *et al.* created unified definitions for four distinct categories of play from an in–depth literature review and created twelve levels of sophistication spanning those categories (see Table 1). The four categories used in the object play coding scale are exploratory, relational, functional, and symbolic play and are briefly defined below:

- **Exploratory Play:** Any child's action upon a *single object* that results from a visually–guided reach and helps provide information about the object or environment. No functional relations exist between action and objects. Examples include: (Level 1) grasping, rubbing, shaking, scratching, banging, poking, mouthing, (Level 2) rolling a car, pushing a button, rocking a horse, and opening/closing doors.

- **Relational Play:** When *two or more objects* are used in combination with each other but are associated without regard to the functions or attributes of the objects. Examples include: (Level 3) pushing apart pop-beads, removing lids from containers, (Level 4) stacking blocks, detaching puzzles pieces, and scooping/pouring objects.

- **Functional Play:** Any conventional use of an object influenced by cultural properties of the object and simple pretend play actions. Examples include: (Level 5) placing a lid on a pot, dumping objects from a truck, (Level 6) drinking from an empty cup, and raising a phone to an ear to talk to a pretend friend.

- **Symbolic Play:** Any scheme in a continuum of play schemes that incorporates items, attributes, contexts not actually present, or the substitution of objects. Examples

include: (Level 9) using a block as a car, or banana as a phone, (Level 10) using figures to load objects into truck, propping a bottle in a doll's arms to feed her, (Level 11) pretending a doll is crying, or claiming a toy stove is hot to the touch.

Table 1: Sequence and definitions of categories used in object play coding scale (from Baranek *et al.* [7])

Category and Level	Definitions	Examples
Exploration of Objects in Play		
Level 1: Indiscriminate actions (2–10 months)	Does not account for functional characteristics; physically manipulates object in unsophisticated ways; treats all objects alike	Tactile manipulations; grasping, rubbing,shaking, scratching, banging, poking, mouthing
Level 2: Simple manipulations of single objects (2–10 months)	Preserves physical or conventional characteristics; discriminates through guided manipulation	Rolling a car; pushing a button; riding a rocking horse; opening/shutting a door
Relational Use of Objects in Play		
Level 3: Takes combinations of objects apart (10–18 months)	Related objects are separated or taken apart	Pulling pop beads apart; taking a lid off a container
Level 4: Presentation/general combinations (10–18 months)	Relating objects by putting them together; combining objects not according to their presentation	Stacking blocks; putting pieces into a puzzle; scooping/pouring
Functional/Conventional Use of Objects in Play		
Level 5: Object–directed (12–18 months)	Actions are directed toward an object	Placing a lid on a pot; dumping objects from a truck
Level 6: Self–directed (12–18 months)	Familiar actions are directed towards the self	Drinking from an empty cup; raising phone to ear and vocalizing
Level 7: Doll–directed (12–18 months)	Familiar actions are directed toward doll figures; child is the agent of the activity	Feeding a doll with a spoon; combing the doll's hair
Level 8: Other–directed (12–18 months)	Familiar actions are directed toward other persons; child is the agent of the activity	Extending a teacup to a person's lips, or a telephone receiver to a person's ear
Symbolic Use of Objects in Play		
Level 9: Object substitution (18–30 months)	Child represents or substitutes one object for another	Substituting a block for a car or a banana as a telephone
Level 10: Agent play (18–30 months)	Child moves doll figures as if they are capable of action	Moving a figure to load blocks onto a truck; propping a bottle in a doll's arms to feed
Level 11: Imaginary play (18–30 months)	Properties are assigned to objects as if they are real; Involves an imaginary object in play or references an object as if it were present	Claiming a toy stove is "hot"; pretending a doll is crying

After the video was coded and analyzed, it was determined that no child, neither those developing typically nor those with an atypical diagnosis, (between the ages of 9 – 12 months) achieved a level of play more sophisticated than functional object directed or self directed play (levels 5 & 6). Regardless of current diagnosis, the children spent a total of 25% of the time engaged in object play with 84.4% of the children engaging in only indiscriminate actions and simple manipulations (levels 1 & 2). Table 2 shows average level of play ability

Table 2: Average level of play ability reached in infants 9–12 months of age (from Baranek *et al.* [7])

Diagnosis Group	Mean Level	Standard Deviation
autism	2.18	.98
other developmental delay	1.70	.67
typical	2.55	1.51

reached per diagnosis group. Baranek *et al.* concluded that observing exploratory play does not help distinguish autism for this age group; however, only typical children demonstrated play ability higher than general object combinations (level 4) reaching functional object directed play or self directed play (level 5 & level 6).

In discussion, Baranek *et al.* raise future questions they would like to explore that are summarized as:

- How often does a child play with a specific toy?
- How often is a specific toy chosen over other toys?
- What type of play is being engaged using specific toys?
- How often is that type of play engaged?

The technology I present in this thesis is designed to help automatically address the questions above. Furthermore, because Baranek *et al.* found that exploratory play alone is not sufficient to distinguish typical from atypical development, my system will focus on identifying exploratory, relational, and functional object play. This will be discussed with more detail in Chapter 4 and Chapter 6. In addition, Baranek *et al.* noted that they had an average interrater reliability of 87%. Achieving this level of reliability required a very fine–grained coding scheme. One of the contributions of the *Child'sPlay* system is that its performance is deterministic and it will label data consistently.

2.1.2 Communication Play Protocol

The Communication Play Protocol (CPP) is a protocol designed to gather a sample of mother–child communication using semi-structured conditions with children ages 18 – 30 months [1, 2]. The CPP focuses on four communicative functions and produces samples

of how mother and child negotiate while interacting socially, requesting items, commenting on items, and narrating between each other with shared objects. The CPP is conducted in a staged play room and is based around the concept that the mother is a "supporting actress" to the child. The mother and child participate in a series of short scenes where the mother is given one cue card per scene to help suggest ways to play with the child. The card describes the plot of the scene, potential props with which to play, and a general direction to try and focus play. The cards do not provide a direct script but provide enough cues to allow the mother to play spontaneously for approximately five minutes.

I will use a method similar to the CPP to elicit samples of specific play behaviors when gathering data from subjects (see Section 5.3 and Section 6.2). I will not use the CPP directly in my studies. Providing general cues to adults will help them play creatively while still ensuring enough samples are collected to build a robust recognition system.

2.2 Capture and Access Systems for Retrospective Analysis

Work by Kientz [39, 36, 37] has focused on embedded capture and access systems to support decision making processes of caregivers for children. Of particular relevance to this work is her system KidCam that supports the tracking of developmental milestones. Her system is enabled by a technology pioneered by Hayes *et al.* [27], known as experience buffers, which Hayes explored in the CareLog system.

2.2.1 Kidcam

KidCam is a prototype system designed to study the use of computer technology to support the early detection of children with special needs [37]. Kientz evaluated the ability of KidCam to support parents and pediatricians in the decision-making process to assess if a child was developing typically over a 4-month period. Her technology consisted of KidCam, a computer supported baby monitor, and companion desktop software that allows parents to collect pictures and videos of their child while also providing age-appropriate prompts for parents to enter developmental health-related information about their child.[1] The software

[1]Prompts were based on the Ages at Stages Questionnaires® (ASQ) [10]

allows review of the child's progress over time at varying levels of detail. If a child has gone too long without completing a specific milestone, the system will alert the parent and add it to a list of questions they can print and bring to their pediatrician at their next scheduled visit. The desktop software also supports the generation of memorabilia such as online video sharing and newsletter–style updates that can be sent to family members.

The KidCam baby monitor is implemented on a Sony Viao-U handtop computer and uses its integrated camera to constantly maintain a temporary buffer of the last 25 minutes of video data. When parents or caregivers observe something they wish to record, they can trigger the baby monitor to save video clips of what just happened by tapping a button on the screen. This experience buffer is similar to those used by the CareLog system [26]. Kientz deployed the full technology in four homes to determine if this computing technology could help increase and encourage the record keeping practices of new parents. A modified version of the desktop software, that does not prompt parents, was simultaneously deployed in four different households for comparative purposes.

This KidCam technology provided increased technological support with the manual tracking of children's developmental progress. Kientz found that although the parents with KidCam recorded more videos, use of the system was still low. In this thesis, I am exploring automated methods for collecting developmental data that can be incorporated into the KidCam smart baby monitor software and that could potentially support automated triggering for the experience buffers.

2.2.2 Lena: Language ENvironment Analysis

LENA is a commercial system designed to help monitor language development in children, from new born to four years old [42]. LENA monitors and measures a child's linguistic progress and their language environment by automatically monitoring child vocalizations, words spoken to the child, conversational turn taking, meaningful speech, and exposure to environmental language. LENA provides frequency and duration information for vocalizations and is designed to help reduce transcription time of audio data for researchers.

Similar to *Child'sPlay*, this system is targeted at early identification of developmental

delays and is designed to generate quantitative statistics about developmental progress. Furthermore, both systems can help reduce data transcription by automatically providing labels of timestamped data for later review. *Child'sPlay* differs from LENA in that it has been designed to monitor development progress associated with play (such as cognitive and motor skills). It should be noted that the augmented toys of the *Child'sPlay* system have the sensing capability to monitor babbles and speech that occur during play – though their audio capability will not be explored as part of this thesis.

2.2.3 Automatic Content Analysis for Social Game Retrieval

Observation of social games between parent and child, such as peak-a-boo and patty-cake, can be important in the early detection of developmental delays. When studying an infant's social ability in research studies, psychologists assess a child's behaviors using recorded videos, such as home movies (similar to the methods described for object play analysis in Section 2.1.1). While computers currently assist in the video-based behavior assessment, it is a manual process where researchers must search for relevant behaviors and score them. This procedure is very time consuming and labor intensive. Work by Wang *et al.* [68, 69, 70] focuses on developing computer vision techniques to automate video filtering and behavior coding of parent–infant social games. In particular the goal of her work is to develop computer vision algorithms to automatically detect and classify social games from unstructured videos. Similar to the goals of this thesis, algorithms by Wang *et al.* are intended to help automate the behavior coding process by enumerating the types of social games a child can play and their frequency, as well as generate other important statistics.

Wang *et al.* has developed an unsupervised algorithm for extracting *quasi–periodic events* from unstructured video by mining for patterns among histograms of visual words [68]. Quasi–periodic events are events which repeat within a specified period of time, but also allow for slight variations within the repetition to occur. For the purposes of modeling and retrieving social games, Wang *et al.* defines social games in terms of these quasi–periodic events as *repetitions of the dyadic interactions, with a range of permissible variations* [70]. By this definition, two individuals engaged in a repetitive interaction, such as patty–cake,

classify as a social game. However, using just the unsupervised method, a child that is repeatedly removing toys from a chest would be classified as a social game. To better classify social games within the video footage extracted with the quasi–periodic algorthm, she then applies support vector machines (SVM [11]) to categorize the video segments according to the type of social game that are present (if any).

In her work on social game detection, she collected two video data sets. The first set consists of three types of social games: patty–cake; rolling a ball back and forth; and tossing a ball back and forth. The data set was collected from ten adults (five dyads) and consists of approximately 40 minutes of footage. Training an SVM classifier (using $\frac{2}{3}$ of the data set for training) on the patterns of visual words, she achieves an accuracy of 94.44% over 18 pattycake sequences, 81.25% over 16 toss-the-ball sequences, and 92.31% over 13 roll-the-ball sequences.

In addition to the adult–only data set, Wang *et al.* collected a second data set consisting of 85 minutes of three parent–child dyads playing freely in a laboratory setting. This data set includes the three games found in the first data set as well as other games. When applying the SVM classifier trained using the adult data set to this second data set, an average recognition rate of recognition rate of 61.41% is achieved.

Wang *et al.* developed her methods concurrently to the development of the *Child'sPlay* system. It should be noted, that the second data set was collected in the same setting as the data sets collected in this theses. In fact, the augmented toys from the *Child'sPlay* system were used during the collection of Wang *et al.* second data set. Similar to Wang *et al.*, I will use SVM to classify object play activities and will apply models trained on adult play to classify play among younger children. However, unlike Wang *et al.*, the *Child'sPlay* system does not make use of computer vision nor unsupervised methods. The combination of computer vision and augmented toys is a logical next step and will be discussed further in Section 8.

2.3 Pattern Recognition for Activities of Daily Living

Many applications in ubiquitous and wearable computing support the collection and automatic identification of daily activities using on-body sensing [34, 64, 46, 43, 54]. In 2004, Bao and Intille showed that the use of two accelerometers positioned at the waist and upper arm were sufficient to recognize 20 distinct household activities, such as brushing teeth or traversing up stairs, using the C4.5 decision tree learning algorithm with overall accuracy rates of 84% [6]. Lester *et al.* reduced the number of sensor locations to one on-body position by incorporating multiple sensor modalities into a single device [44]. Using a hybrid of both discriminative and generative machine learning methods (modified AdaBoost and HMMs to smooth the results), they recognized 10 activities of daily living with an overall accuracy of 95%.

In each of these works, the sensors remained on-body in a fixed orientation and often in a fixed position limiting the degrees of freedom experienced by the sensor, hence limiting the parameters that must be learned. The *Child'sPlay* system, however, can have an activity performed with the sensors in any number of orientations. This increase in parameter space could have dramatic effects on both training time, the number of examples required, and as a result, the classifiers that can be learned (see Appendix D.3). For reasons of practicality, the training time of the system should not exceed the time it takes to manually annotate the behaviors being studied. In this case, each question posed could potentially require additional model training and be analogous to psychologists recoding the data by hand to address different research agendas.

2.4 Evaluation of Continuous Activity Recognition Systems

Recently, a method was developed for visually representing performance that explicitly accounts for the various types of recognition errors that an automated system can incur, known as Multiclass Segment Error Table (MSET) [71, 72]. MSETs represents the total duration of the data as a rectangle and subdivides it into sections corresponding to the different classification results. This method provides a relatively complete picture of the overall performance and error distribution for the system. Multiple recognition methods

can be compared at a glance by aligning the corresponding diagrams and visually comparing the relative area of the true positive division and relevant error divisions.

Section 3.2 shows modification to the MSET framework that accounts for a more diverse range of error distribution when evaluating recognition systems. This advanced representation could be used as a method to evaluate the performance of various algorithms applied to identify object play behaviors.

2.5 Automating Cognitive Assessments using Tangible Interfaces

A Graspable user interface, according to Fitzmaurice, *"provides users concurrent access to multiple, specialized input devices which can serve as dedicated physical interface widgets, affording physical manipulation and spatial arrangements"*[20]. Ullmer and Ishii state that *"Generally graspable and tangible interfaces are systems relating to the use of physical artifacts as representations and controls for digital information"* [66]. Graspable and tangible interfaces are used in a variety of applications including (but not limited to) evaluations of construction tasks, edutainment, interactive toys, creative play systems, and structural design systems. Of particular interest to this work is the use of tangible interfaces as an aide for clinical assessment.

The assessment of cognitive abilities is an important aspect of evaluating a child's developmental progress. The development and retention of cognitive and motor skills can be assessed by observing the constructional ability of an individual. Constructional ability can be quantified by observing performance on drawing, assembly, and building tasks. For example, a common 3D construction task is to replicate a specific spatial structure representation with building blocks. The completion of these construction tasks requires the ability to perceive the target shape, reason about the spatial structure of the shape, develop a plan to construct the shape, and physically build the shape [45, 24].

Typically, assessment of construction ability is performed manually in a clinical setting by highly trained specialists. Common metrics, such as task completion time, accuracy of construction, assembly order, and analysis of construction strategy, are subjective and become more difficult to assess as shape complexity increases. The manual scoring of these

tasks often introduces bias to the assessment and may decrease test reliability [24]. As such, recent research has focused on automating clinical assessments. In particular, research in graspable and tangible user interfaces has explored automating the clinical assessment of 3D cognitive construction.

Cognitive Cubes is a prototype system designed to automate the clinical assessment of cognitive construction tasks [62]. The Cognitive Cube system consists of a tangible user interface and video projection system. During a session an administrator selects a predefined target shape that is projected onto a screen. The participant reconstructs the projected, rotating shape using ActiveCube, a tangible user interface for describing three-dimensional shapes [40].

The Cognitive Cubes system analyzes the data collected from the ActiveCubes, offline, after a construction task is completed. The system computes four assessment measures based on the similarity of the participant's built object to the target object: the similarity at time of completion, the duration of the construction task, the rate of completion, and the consistency of progress.

A pilot study and two in–depth studies were conducted comparing the Cognitive Cube system to manual assessments. Forty-three participants ranging in age from 22–86 participated in the studies. Two of the participants had mild Alzheimer's disease. The studies showed that Cognitive Cubes is sensitive to cognitive factors, increased scoring measurement resolution, and increased reliability of assessment when compared to manual scoring of 3D construction tasks. Strengths of the system include consistency of administration, and sensitivity to cognitive deficiencies by recording data on the often ignored intermediary steps of construction.

Similar to the Cognitive Cube system, my system aims to increase the consistency of manual annotation and provide the ability to record relevant data that would have otherwise gone unnoticed by human observers for object play behaviors. However, unlike the Cognitive Cube system, *Child'sPlay* will target a much younger population and consists of wireless components.

CHAPTER III

PREVIOUS WORK: ACTIVITY RECOGNITION TECHNIQUES FOR CONTINUOUS, MOBILE WIRELESS SENSING

Activity recognition is the problem of detecting and identifying activities in time-varying sensor data [50]. As mentioned in the previous chapter, there have been several projects involving the automatic recognition of daily activities as recorded by wireless sensors. This chapter briefly describes one of my previous projects involving mobile wireless sensing. While the system described below is not directly related to the identification of object play activities, it was a prototype system designed to automatically recognize activities and index them for later review. After, I will discuss previous, collaborative work involving the types of errors that can occur during continuous recognition and how Error Division Diagrams (EDDs) can be used to help researchers visually compare the performance of recognition systems in terms of these errors. Based on this analysis researchers can select the system that best suits their needs [50]. This chapter closes with a discussion of the impact that different recognition error types might have on an intelligent interface designed for retrospective review of object play.

3.1 Classification using HMMs and the Georgia Tech Gesture Toolkit

In this section I present my initial work investigating the use of wireless sensors to assist in the naturalistic observation and care of children with autism. In particular, I describe an on–body system that provides continuous recognition of mimicked autistic self–stimulatory behaviors using three wireless accelerometers [77]. This pilot study provided a proof-of-concept system that is capable of collecting data from a child with autism and can also automatically provide indices into that data to highlight the self–stimulatory behaviors for later review. Our initial results are computed using a neurotypical adult and indicate that an automatic indexing system for self-stimulatory activity is feasible. However, there are many practical issues that may make a wearable system for the target population of children

difficult to deploy [38, 21]. This section also discusses the recognition components of the Georgia Tech Gesture Toolkit that were used in this project, followed by a brief discussion of the implications for recognizing object play behaviors in toddlers.

3.1.1 Wireless On–body Sensing to Support Children with Autism

Autism is a developmental disorder affecting a child's social development and ability to communicate. Children with autism will often exhibit behaviors such as vocal stutters and brief bouts of vigorous activity (*e.g.,* violently striking the back of the hands) to cope with everyday life. Depending on the child's level of functioning, these highly individualized, self-stimulatory ("stimming") behaviors can be disruptive, socially awkward, and even harmful. Caregivers and researchers would like to explore the correlation between these stimming behaviors and environmental factors, behavioral treatments, mood, and other physiological markers.

To assist in this analysis, I aimed to automate the recording and analysis of these behaviors [77]. Although it is impractical for a researcher to monitor a given child continuously for episodes of stimming, an intelligent monitoring system could collect daily data from the child and filter it so that only the stimming episodes are highlighted. An automated data collection system may provide insight into a given child's mental and physiological state. It may also provide detailed, quantitative data for researchers in the field, which is currently rare.

The initial results indicate that an automatic indexing system for stimming activity is feasible. Our data set consists of acceleration data generated from a neurotypical adult mimicking autistic stimming behaviors while performing unscripted activities. The accelerometers were positioned on the right wrist, the back of the waist, and the left–ankle. Seven stimming behaviors and intermediary unconstrained "non-stimming" activities were modeled using hidden Markov models (HMMs) via the Georgia Tech Gesture Toolkit (GT^2k) [73] (see Section 3.1.2). I explored the performance of these models in both isolated and continuous settings. The isolated HMM experiments assumed slight noise in data segmentation and achieved accuracy rates of 91.0 percent. In the continuous recognition experiments,

exact segmentation of the stimming events was not possible due to minor insertion errors. These fragmentation errors (rapid alternation of classes at the boundaries) produce an overall system accuracy of 68.6 percent. However, I improved segmentation accuracy by using insertion penalties and smoothing during the model alignment process. I achieve a recall rate of 100 percent for the self-stimulatory events with 92.9 percent precision including identification of non-self-stimulatory activities.

3.1.2 Georgia Tech Gesture Toolkit: GT^{2}k

In 2003 I developed and released The Georgia Tech Gesture Toolkit (GT^{2}k) [73]. The GT^{2}k provides a publicly available toolkit, which leverages Cambridge University's speech recognition toolkit HTK, for developing gesture–based recognition systems [29]. Since its release, the GT^{2}k has been used in over 100 projects across 3 continents to provide tools that support gesture recognition research. It has also resulted in a secondary, more accessible toolkit, GART: the Gesture and Activity Recognition Toolkit [47].

GT^{2}k provides a user with tools for preparation, training, validation, and recognition using HMMs for gesture–based applications (see Appendix C for mathematical details). In the simplest case, recognition can be performed on one gesture at a time. This technique is known as *isolated* gesture recognition. However, the more practical use for my purpose is to perform *continuous* recognition on a sequence of gestures within a contiguous block of data. Knowledge of the possible sequences of gestures can be presented to GT^{2}k in the form of a rule–based or stochastic grammar. Grammars allow GT^{2}k to leverage knowledge about the structure of the data, which aids in continuous recognition by constraining the gesture classification with respect to the previously classified gestures. Grammars also allow users to define complex gestures as a sequence of simpler gestures.

3.1.3 Implications

GT^{2}k provides a flexible and powerful framework for using HMMs in continuous activity recognition systems. Of particular interest to this thesis is the support for bi–gram, tri–gram, and N–gram grammars that will allow greater representational power for expressing object-play behaviors in young children. For the purposes of recognizing children's object

play activities, there is an implicit structure for simple exploratory behaviors in so much that a child must pick up a toy before he can shake the toy. For more complex relational play actions, such as stacking and unstacking, grammars can be used to remember object state. However, specifying a rule–based grammar for every combination of toy and object play activity could be tedious. For this reason, I will explore the use of stochastic grammars to allow domain knowledge to influence model alignment.

It should be noted that whatever the combination of sensors, algorithms, and accuracy, the viability of the end solution is determined by the influence of the recognition errors on the retrospective review task. There are certain error types that can be ignored or overlooked with respect to the retrospective review of object play behaviors. However, a discussion of acceptable and unacceptable error types is delayed until Section 3.3, after the introduction of work describing the different error types that can occur during continuous recognition.

3.2 Quantitative Evaluation Metrics for Systems Supporting Retrospective Analysis

In this section I will discuss my collaborative work involving the description of error types that can occur during continuous recognition and how Error Division Diagrams (EDDs) can help researchers compare the performance of recognition systems visually to select the system that best suits their needs [50]. I will also discuss the impact these errors might have on an intelligent interface designed for retrospective review of object play.

3.2.1 Disadvantages of a Single, Numerical Metric

Standard accuracy metrics do not always account for the impact that different error types have on applications. For example, in automatic recognition systems, a trade–off exists between identifying all instances of an activity and obtaining accurate event boundaries. These trade–offs have different impacts based on how the recognition technology is being used. For example, when coding videos for play, some researchers may be interested in the exact duration of play events (*e.g.,* how long a child rocked a wobbly toy back–and–forth). Other researchers may only be interested in the number of times the play event

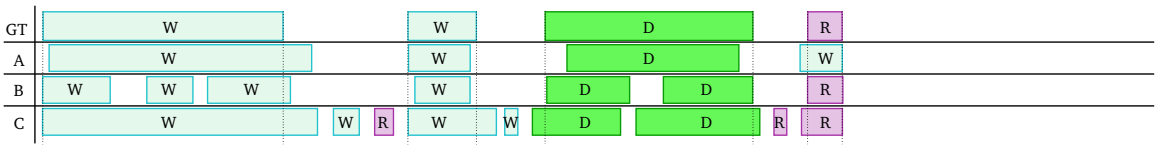


Figure 2: Sample hypothetical ground truth (GT) labels for a simple domain that includes waving (W), dropping (D), and rolling (R) toys, along with the hypothetical predicted labels for three different recognition systems (A, B, and C) that yield equivalent accuracy.

occurred, while others may merely want to know if the event occurred within a segment of video. When considering these different types of usage scenarios, a single numerical metric can often be misleading. For example, Figure 2 shows a recognition task along with the hypothetical predicted output of three systems. The top row consists of ground truth events while the subsequent rows consist of predicted output values. All three systems yield the same frame–level accuracy of 66 %. However, as illustrated, the three systems do not have identical output. In fact, each system produces different types of frame and event level errors. Appendix B.1 illustrates and describes frame, event, and segment analysis in more detail. The impact of these errors may vary in significance depending on the application as well as the level of analysis being preformed.

3.2.2 Types of Errors Encountered in Continuous Recognition

There are many types of errors that can occur during continuous recognition involving correspondence issues between activity boundaries and labels. Figure 3 shows the output of nine different recognition systems, $A - J$, where each illustrates a specific error type common to continuous recognition. The definitions and specifics of these error types are listed in Appendix B.3.

3.2.3 Error Division Diagrams

Error Division Diagrams (EDDs) are a way to represent graphically the overall performance of a recognition system including both the distribution of errors (according to type) and the percentage of null activity present. These rectangular diagrams organize errors according to type and severity by representing each error type as a corresponding percentage of the entire column. Figure 4(a) represents two EDDs with symbolic labels for illustrative purposes.

Starting from top to bottom, these labels represent the percentage of frames that were true positives, true negatives, overfills, underfills, fragmentations, merges, insertions, deletions, substitution-fragmentations, substitution-merges, and substitutions with the black horizontal line indicating the division between mild and severe errors. Figure 4(b) is the numerical version of Figure 4(a).

The two top divisions of EDDs represent the percentage of true positive and true negative instances recognized, respectively. These two divisions account for all of the data that was correctly identified by the system. All divisions afterwards represent errors which increase in severity with minor errors towards the top and more severe errors at the bottom. For example, overfill (O) and underfill (U), indicate simple boundary errors whereas insertions (I) and deletions (D) of events are more serious classification errors. In Figure 4(b), System A and System B correctly identify the same percentage of events, however, System B has less severe errors.

Figure 4(c) shows an EDD comparison of multiple systems. With these diagrams, recognition methods can be compared by inspecting the percentage of errors below the serious error line and relative area occupied by other errors. For example, the large percentage of area devoted to overfill (O) and underfill (U) indicates that System B has event boundary

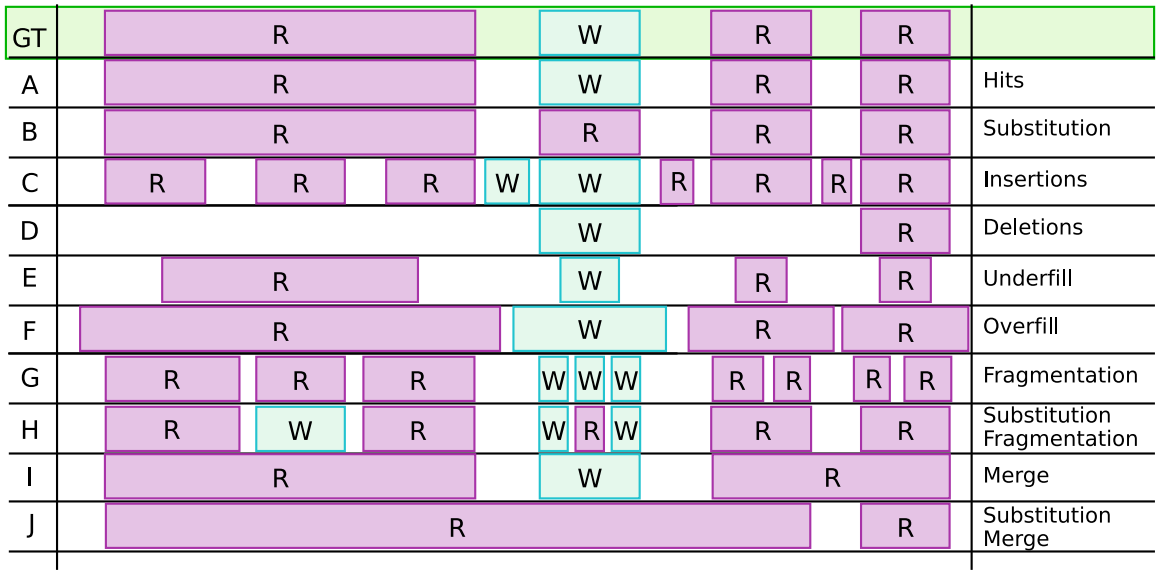


Figure 3: Different boundary and label correspondence error types that can occur in continuous recognition systems. The ground truth labels are highlighted at the top.

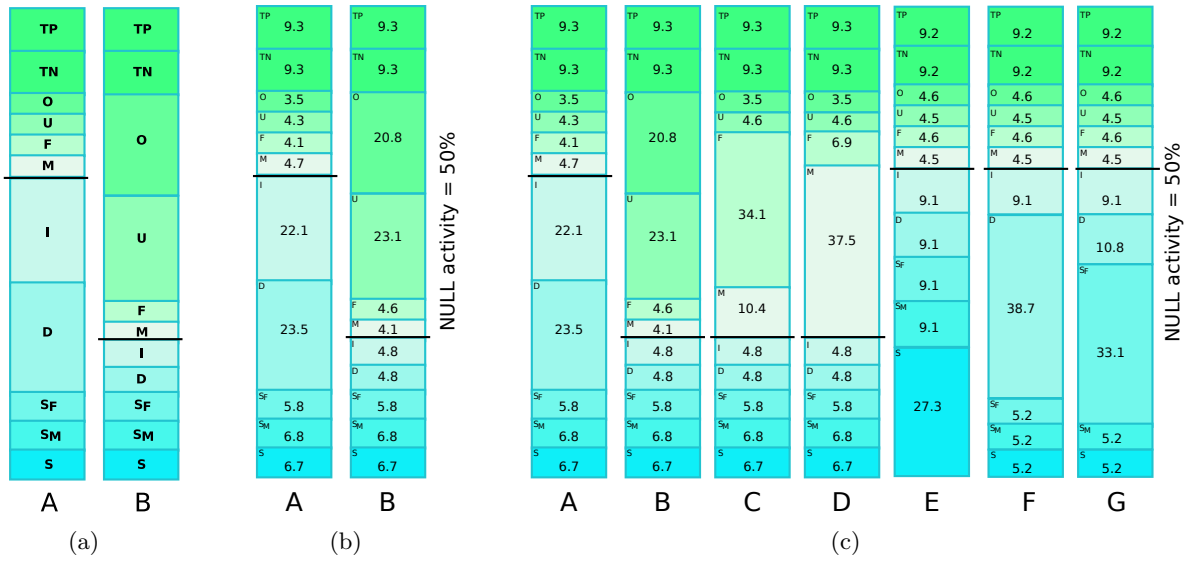


Figure 4: Error Division Diagrams comparing recognition systems. Part (a) and part (b) show identical comparisons, however, Part (a) represents EDDs with symbolic labels while part (b) represents EDDs with numeric labels. Showing, from top to bottom, the percentage of frames that were true positives, true negatives, overfills, underfills, fragmentations, merges, insertions, deletions, substitution-fragmentations, substitution-merges, and substitutions. The dark horizontal bar indicates the division between severe and mild errors. Part (c) shows EDDs comparing multiple systems

errors. Systems C and D have a large percentage of fragmentation and merge errors (respectively), indicating a difficulty in determining the duration of events. Systems E, F, and G, on the other hand, have more serious errors whereby they incorrectly identify events, delete events, or state that events occurred where they did not.

3.3 Implications of Error Types for the Child'sPlay *System*

Recognition does not need to be perfect to be useful for retrospective analysis. For the task of reviewing object play, there are some errors that are less severe than others depending on the task. For example, if the recognition portion of the *Child'sPlay* system produces overfill and underfill errors, it will highlight the occurrence of play events but provide inaccurate timings of the events. If the goal of the retrospective analysis is to count event frequencies, overfill and underfill are not serious errors. However, if the analysis goal is to tally the duration of specific events, these errors would be slightly more serious – the identified occurrence of the event can guide the user to the temporal location, but it would require

the user to identify the boundaries of the event. Deletion and insertion errors, if numerous, could have a drastic impact on both frequency and duration counts. If these error rates are high, the user could potentially waste time dismissing false positives or searching for missed false negatives.

By the same token, if the goal of the *Child'sPlay* system is to help identify the achievement and maintenance of specific levels of object play sophistication, the deletion and insertion errors may be less serious. In the case of achievement, the system need only to identify a small portion of instances for a specific behavior. Not all events need to be identified to prove a developmental goal has been reached. Even if deletion errors were frequent, the system need only recognize one instance to prove achievement. Likewise, in order to verify the maintenance of a skill, only a fraction of instances need to be identified. This justification partially holds for insertion errors as well. For example, the *Child'sPlay* system could falsely identify multiple instance of relational play. However, if even a fraction of the events are true positives, (which the user can screen for correctness) then the system has shown achievement of relational play. If, however, the system failed to identify any true instances of relational play (zero true positives, 0% recall), the insertions would be misleading. While EDDs are not used directly in the *Child'sPlay* system, the discussion above helps highlight which types of recognition errors would be most detrimental when using the *Child'sPlay* system to analyze developmental progress.

CHAPTER IV

AUGMENTED TOY DESIGN

This chapter discusses the motivations behind the selection of the activities supported by *Child'sPlay*, followed by a discussion of the sensing challenges and requirements that must be met by the toys. I end the chapter with a brief discussion on the initial play tests with children as well as the implications for the *Child'sPlay* system and subsequently collected data sets.

4.1 Activities to Recognize

Child'sPlay will support a subset of play activities similar to those studied in clinical research. In particular, the system will automatically generate quantitative data from observations of children engaged in object play similar to that produced by the coding scheme of Baranek *et al.* [7]. These measures include the frequency with which an object is played, the time spent attending between different objects, and the highest level of play sophistication reached by a child. Based on the scale produced by Baranek *et al.*, as well as conversations with developmental psychologists [59, 4], our technology will focus on recognizing toy manipulations such as grasping, exploring, shaking, rolling objects, pulling apart Lego^{TM} Quatro–compatible blocks, assembling, pouring, stacking blocks, nesting objects and early imaginary actions (see Table 3). These actions form the basis of more complicated levels of play whose recognition is an important first step towards identifying more complicated play structures, such as symbolic play (the recognition of which is beyond the scope of this thesis). Toys and activities that appear in our pilot data sets (discussed in Section 4.4 and Section 5.3) differ slightly to those that appear in our final adult and child data sets (discussed in Section 4.5).

Table 3: Elementary levels of object play along with canonical examples [7]

Category	Levels	Examples
Exploratory	L1: indiscriminate actions L2: simple manipulation	grasp, rub, shake, bang, mouth rolling toys, pushing a button
Relational	L3: takes combinations apart L4: general combinations	pull apart assembled toys, remove lids stacking, scooping, pouring
Functional	L5: object directed L6: self directed	covering with lids, dump payloads imaginary drinking or talking on phone

4.2 Sensing Considerations

Several trade-offs exist in the development of a play sensing system, including sensor type, power consumed, and form factor. The types of sensors used and form factor of the toys influence the quality of data that can be recorded. Regarding the design requirements of the form factor, the sensor should be easy to charge, have maximum protection from daily use, be unobtrusive, and remain in position during use. Embedding the sensor within the toy addresses many of these issues and maintains the original safety properties of the toys. It can also help keep the sensor in the proper position to allow for consistent data recording and can prevent the sensor from becoming exposed to the child while in use. The form factor also determines the ease with which the sensors can be accessed by caregivers to remove for toy cleaning maintenance and for charging the battery. However, finding a balance between ease of access for adults and preventing the children from accessing the sensors can be difficult. Designs that require manual dexterity, such as screw tops and/or constant force are often good for preventing children from accessing the hardware.

While no one sensor is ideal for automatic play recognition, a fusion of sensors can help increase the range of activities that can be detected. Our toy designs favor the multiple modality BlueSense integrated wireless sensor package [56]. The BlueSense sensors detect motion, sound, and touch via two audio analog inputs, two capacitive touch-sensing inputs, and an on–board 3–axis accelerometer. They measure about 1.8x1.8 inches (4.6 cm) and can transmit data continuously for about 10 to 12 hours using a light rechargeable 3.6V 750mA battery.

The sensing modalities supported by BlueSense are well suited to the range of play

Table 4: The *Child'sPlay* augmented toys and the activities they promote. The groupings correspond to different early levels of object play similar to those described by Baranek *et al.*

	Exploration of Objects in Play									Relational Use of Objects in Play			Functional Use of Objects	
	Indiscriminate Actions					Manipulations of Single Objects				Takes Combinations Apart	Presentation and General Combinations		Object Directed	Self Directed
Toys	**explore**	**grasp**	**shake**	**rub**	**bang**	**rock**	**spin**	**slide**	**roll**	**take apart**	**stack**	**assemble**	**wears**	**runs towards**
Ring	✓	✓	✓		✓		✓	✓	✓		✓		✓	
Block	✓	✓	✓	✓	✓			✓			✓			
Caterpillar	✓	✓	✓	✓	✓			✓	✓	✓		✓		✓
Domes	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	
Legos	✓	✓	✓		✓			✓		✓	✓	✓		
Lid	✓	✓	✓		✓		✓	✓	✓	✓	✓		✓	
Puppy	✓	✓	✓	✓	✓			✓						✓

activities that *Child'sPlay* system is to detect. However, there are two main disadvantages to using an integrated approach. First, because I am using a smaller, lighter battery, it will require more frequent charging. Second, using a centrally embedded sensor package means that some of the integrated sensors will not have optimal positioning. For example, the microphone will not have external exposure, as it is located inside the toy. Likewise, the centralized sensor location can cause issues for capacitive sensing as well. Two of our designs explore using conductive threads and fabrics to address this issue for sensing touch. These are the plush cube and plush caterpillar designs[1]. These designs, as well as the other toys in *Child'sPlay* will be discuss in the next section, Section 4.3.

4.3 Toy Selection and Form Factors

I have designed and implemented seven toys to collect data about toddler–object play behaviors. These toys include a plush puppy rattle, a plush caterpillar, a plush cube, plastic Lego^{TM} Quatro compatible blocks, a plastic ring stacking toy compatible with the Fisher–Price^{TM} Rock-a-Stack toy, an abstract shape resembling a cooking pot lid, and two plastic dome toys compatible with the Fisher–Price^{TM} Stack-&-Roll Cups toy (see Figure 6). All of the toys are designed to use the BlueSense sensing unit enclosed in a friction-fit plastic case that is embedded within the toy. The plush puppy rattle and plush caterpillar toys are adorned with smiling faces, to encourage social engagement with the toys [59]. The

[1]The plush caterpillar was designed and implemented in collaboration with Wooyoung Sung, Pamela Griffith, Michael Genovese, and Scott Gilliland as part of a project for the Mobile and Ubiquitous Computing class in Fall 2007. Details on the plush cube design can be found in Appendix F.

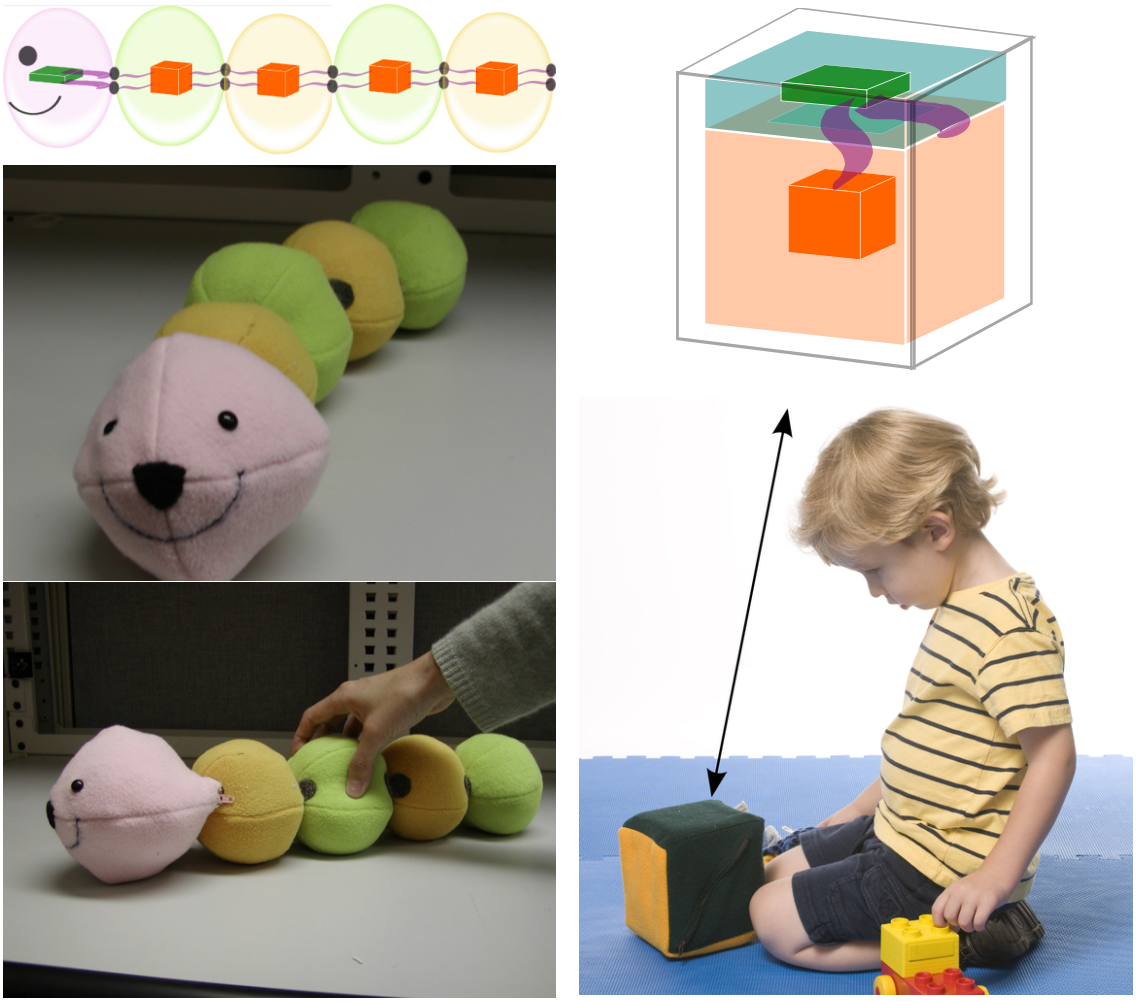


Figure 5: Plush toys designed to detect touch via capacitive sensing

ring, lid, and dome toys are rounded objects based on a similar circumference to encourage stacking, covering, and scooping activities. Table 4 lists the prototype toys and the object play actions they promote according to level of sophistication. An important specification of our design shared by all toys was safety. All toys are large enough so children cannot swallow them. All of the toys except for the plush caterpillar and the cooking pot lid are modeled from existing toys approved for infant use. Each toy continuously transmits data via the Bluetooth sensor to a mobile computing device where it records and processes the data. Example data collection platforms used are the Sony Vaio UX10 and an IBM X31 laptop. Data from a single toy has also been collected on a Nokia cell phone platform.

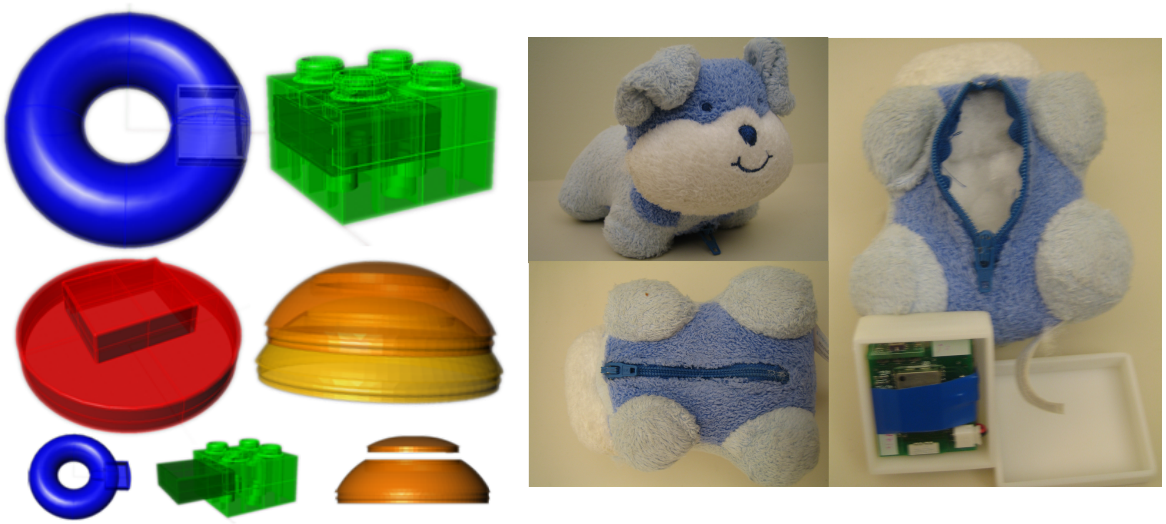


Figure 6: Left: CAD models of the plastic toys used with *Child'sPlay*. Right: the plush puppy rattle with sensing unit exposed.

4.4 Pilot Play Sessions with Initial Toy Designs

I have conducted initial play tests with the seven toys. These tests included three toddlers (and respective parents) over a minimum of two 20 minute sessions per participant where each session resulted in at least five minutes of play time with some of the augmented toys. These informal sessions, which involved free play, allowed me to test the durability, appeal to children, and data transmission capability of the toys.

From this play pilot, I learned that the ABS plastic and conductive textile toys were durable, functional as toys, of interest to children, and concealed the sensor from the participants. The toys withstood throw, drops, and kicks that occurred during play. However, the design of the plush caterpillar proved to be flawed. The design of the caterpillar was such that the sensor was positioned in the head of the caterpillar with the remaining three body segments connected via conductive Velcro® (see Figure 5). The segment connections proved too brittle and detached unexpectedly during play. Furthermore, the body segments were stuffed with a soft cotton that did not retain shape when gripped. This caused baseline issues for the capacitance sensing as the segments did not return to their original shape when released. The caterpillar toy will not be used in subsequent studies due to the difficulty to correct these flaws versus the benefit of including the toy. The plush cube,

which was designed with conductive materials and a stiffer foam core, retains it shape when released and does not suffer from similar baseline issues.

In addition to the flaws with the plush caterpillar design, the play tests did expose an important challenge for constructing the automatic recognition portion of *Child'sPlay*. The most prominent challenge is collecting enough training examples to build models for recognition. While I can script play scenarios to help encourage children to engage in the types of play the *Child'sPlay* system is trying to detect, there is no guarantee that the child will be willing or able to comply. This problem is further complicated by the fact that our target age group is children ages 10–24 months old. Some of these children may not have yet formed language, nor will they be receptive to instructions by adults. Thus, I cannot instruct them to play with the toys in a way that will be guaranteed to elicit proper training data. One possible solution to this problem is to bootstrap the models with data collected from adults engaged in semi–scripted play. Chapter 5 discusses the collection of a pilot data set using adult subjects and my initial results when detecting relevant play activities among the adults using both augmented and regular toys. Chapter 6 describes the collection of a larger play data set including both children and adult play sessions as well as discusses the application of more robust statistical adult models to detect children's play.

4.5 Modifications and Final Toy Designs

Both the toys' form factors and the specific play activities to be recognized were modified after the collection of each pilot data set. The initial toy designs were used in the play tests described in Section 4.4 and in the collection of the pilot adult play data set described in Section 5.3. The initial plastic toys were white in color and designed to be compatible with certain existing off–the–shelf toys. While the augmented toys interacted well with their commercial counterparts, the augmented toys were not explicitly designed to interact with each other. As will be discussed with more detail in Chapter 5, play interactions involving multiple augmented toys are generally easier to detect automatically than play interactions involving a mixture of regular toys and an augmented toy. Hence, by designing toys to interact with commercial counterparts rather than other augmented toys, the initial set of

Figure 7: Final version of the plastic dome toys

toys may have increased the complexity and difficulty of both modeling and recognizing elementary play behaviors.

While interactions between augmented toys and regular toys are likely to occur in natural settings, I feel it is important to leverage the toy designs to maximize the opportunity of recording and identifying higher levels of play. As such, the plastic dome and plastic ring designs were modified to encourage a wider range of relational, functional, and symbolic play within the augmented toys. Color was also added to all the plastic toys to broaden appeal.

4.5.1 Modifications to the Plastic Dome Design

The main modification made to the plastic dome toy is to increase the size of the plastic dome toys both in circumference and curvature such that they can allow the nesting of other augmented toys inside, such as the plush puppy rattle or Lego^{TM} Quatro toys. Although the increase in size renders the plastic dome toys incompatible with the Fisher–Price^{TM} Stack-&-Roll Cups, it still maintains the same functional properties of the original toy. Namely, the plastic dome toys are still able to stack one on top of the other to form a tower and also assemble together, end to end, to form a ball (see Figure 7). The modification of the domes' sizes is important as it allows for other toys to be nested inside the domes, which increases the types of play that can occur while using the domes [4]. For example, a single plastic dome can now be used to hide the plush puppy rattle from sight or used as an imaginary vehicle for the puppy. During initial play tests with the new design, the plush puppy rattle was often rocked to sleep in an inverted dome, flown through the air as if in an airplane, or sledded across the ground. A popular play motion among children

Figure 8: Final version of the plastic ring toy

and adults alike was to conceal the puppy within both domes (forming the ball) and then rolling it around on the ground.

4.5.2 Modifications to the Plastic Ring Design

The main modification made to the plastic ring was to flatten it (making it more disc shaped) and to place recesses in the top and bottom. These indentations allow the smaller red plastic dome to be interlocked with the plastic ring toy but do not allow the larger blue plastic dome toy to fit. Instead the red plastic dome can interlock with either side of the green ring to form a flying saucer shaped toy in which the plush puppy rattle toy can also fit (see Figure 8). When the child tries to assemble the blue plastic dome and the

green plastic ring, relation interactions occur as the child discovers that the toys do not fit. In the process of flattening the plastic ring the circumference of the inner whole was also increased, allowing the plastic ring to be worn on the wrist or leg by the child. As with the modifications made to the plastic dome toys, the modifications made to the plastic ring toy also increase the types of play that can occur while using the plastic ring toy [4]. During the initial play tests with the new design, the plastic ring was often worn around the play space while playing with other toys. It was also frequently used with the plush puppy rattle toy as a feeding dish, a bed, and an acrobatic hoop. When the red plastic dome is assembled with the plastic ring, it is often used as a flying saucer to abduct the puppy. One of the younger children even used this combination of toys as a drum.

CHAPTER V

DETECTION OF PLAY BEHAVIORS WITH ADULTS USING A MIX OF AUGMENTED AND REGULAR TOYS: A PILOT STUDY

When undertaking the development of any recognition system, it is important to have a baseline of how the system would perform under ideal conditions. In the context of a play recognition system, it is beneficial to know how many play activities can be detected and to what degree of accuracy they can be detected. Because our target population consists of infants and toddlers, constructing a structured experiment to obtain such a baseline can be difficult (see Section 5.2 for more details). As such, in late 2007 I conducted a pilot study using adults.

This chapter begins with the research questions and hypothesis that this study addresses (Section 5.1). The chapter then provides a discussion of the feasibility of using adults to obtain a baseline of activities that can be reliably recognized (Section 5.2) followed by a description of the experimental procedure (Section 5.3). After, a detailed description of the data is provided (Section 5.4) along with a description of the applied algorithm (Section 5.5) and associated experimental results (Section 5.6). The chapter concludes with a discussion of the results (Section 5.7) and their implications for subsequent studies with the target population (Section 5.8).

5.1 Research Questions and Hypothesis

This study is designed to address Research Question 1 and the following subquestions:

1. How many distinct activities can be detected?
2. With what level of accuracy can we detect these items?
3. How do user–independent, user–dependent, toy–dependent, and toy–independent models compare in performance?

In particular, I demonstrate that standard techniques for detecting activities in wireless systems will allow for detection of primitive play that achieves rates significantly better than random selection. Accuracy will be lowered by high insertion errors and user–dependent, toy–dependent models will perform best.

5.2 Adults as a Baseline

Using adults in place of children has both advantages and disadvantages. As mentioned in Section 4.4, children often do not do what they are asked or told. Not only does this factor make obtaining a sampling of the activity space to be modeled difficult, it also can make data collection very time consuming and yield comparatively small samples of usable data. During the play tests described in Section 4.4 there were tantrums, bouts of pouting, and wandering about (to explore everything except the augmented toys). These non–play activities resulted in some data collection sessions that lasted over 90 minutes with approximately 15 minutes of playtime using the augmented toys (or in close proximity of the toys). The amount of usable data may be less than the total playtime due to heavy parental influence. For example, one child was playing with two Lego^{TM} Quatro bricks but was unable to separate them. Afraid that the child might side track into a tantrum, the parent quickly separated the blocks for the child. While parental involvement is expected during play, it can reduce the number of examples of the child performing certain activities. To ensure that the pilot was conducted in a reasonable amount of time and that the amount of usable training data was maximized, this study was conducted using adult participants.

5.3 Method

To gather the baseline, five adults were recruited, two female and three male. Each subject participated in a minimum of two sessions, with each session on a different day. Data was collected over the course of seven days with play sessions ranging from 7 – 26 minutes (mean 16.32 minutes, STD 7.17 minutes). Each participant was seated at a table and then presented with a variety toys in an opaque bag (see Figure 9). Participants were asked to shake the bag twice (to provide a means of verifying data synchronization for post processing) and then were asked to retrieve toys from the bag. Some participants chose

to dump the contents of the bag onto the table while others selected a single item from the bag each time the toy was requested. Once the toys were on the table the participants were then instructed to perform a series of play tasks (see Table 5). At the end of each session, the toys were placed back into the bag, and the bag was shaken once again for data synchronization purposes.

Table 5: Primary tasks asked of adults while playing

Prompt Question	**Desired Behavior(s)**
"Let's play knock-knock with . . . ?"	banging
"Let's play leapfrog with . . . ?"	grasping, moving, stacking
"What sound does it make?"	shaking
"Find me the feature / reflection"	exploratory manipulations
"Can puppy play with . . . ?"	imaginary
"Does it fit in there"	relational, banging
"Give me/Find me/Hide"	unintentionals, push, pull
"Let's tower some blocks . . . "	join, separate, stack
"Does it Spin or Roll ?"	bumping, pushing, rolling

The instructions provided during each session were designed to elicit certain play behaviors without directly asking for specific activities to be performed. For example, to gather examples of shaking the toys, the participant was asked, "Find me a toy that rattles." Typically the participant would then pick up each toy, in turn, and shake the toy to determine if the toy produced a noise. Indirect questions were used with the hope of producing a more naturalistic data set. Furthermore, these questions also helped engage the participants and provide suggestions on how they should play. Participants were asked a series of questions during each session to ensure that multiple examples of each play behavior were obtained. It should be mentioned that strict structure was not placed on the play session. Questions were not asked in a specific order and were often adapted to fit the context of play currently seen on the table as participants would often manipulate toys in unexpected ways. To keep the rhythm of play going, questions were often adapted to what the participant did rather than being based on the expected outcome that the question was meant to elicit. This procedure resulted in play sessions that include subsets of activities over various sessions with varying degrees of sophistication. In other words, the same questions were not asked of each participant; therefore, the data sets are not uniform across participants nor trials.

Toy	Type
plastic dome	augmented
plastic ring	augmented
reflective lid	augmented
Lego^{TM} Quatro	augmented
plush puppy rattle	augmented
Snap–Lock Beads	Fisher Price OTS
Stack-&-Roll Cups	Fisher Price OTS
Rock-a-Stack	Fisher Price OTS

Figure 9: Augmented and Off-the-Shelf (OTS) toys used during adult play sessions

5.4 Data Description

The five augmented toys depicted in Figure 9 were used to collect data during each of the play session. Each augmented toy contains a 3–axis accelerometer that samples at approximately 50 Hz. Therefore, each play session produces 15–dimensional data (3 axes per toy) with approximately 49,000 samples per session. A total of 12 play sessions were completed by the 5 participants resulting in 3.8 hours (228.49 minutes and 692,520 samples) of 15–dimensional data. Participants only performed one play session per day.

In addition to collecting accelerometer data, motion jpegs from a single camera were also collected during each play session. These images provided ground truth for data labeling. Both the accelerometer data and image data were collected on the same machine to simplify data synchronization issues during labeling of the ground truth data. Data collection was also confined to a single device, an IBM X31, to act as a prototype for a self–contained device that could be deployed in a household. The data was labeled using specially developed software, *TSview*, which visually aligns the ground truth images with the sensor streams (see Figure 10).

Two students independently labeled the 15–dimensional data using the *TSview* software. These students were trained for approximately 1.5 hours to identify the elementary play activities of interest and for 20 minutes to use the *TSview* software. During training, each

person was provided with a coding manual (see Appendix H) indicating the 24 object-toddler activities to identify within the data (see Table 6). Although the 15–dimensional data stream represented 5 toys, only one label is provided for any one instant in time. Therefore, when identifying each of the 24 actions, the associated toy must also be identified by the label. If it is assumed that no toy interacts with any other toys, then this scheme would lead to 120 distinct classes (24 *actions* $\times$ 5 *toys*). However, if more than one toy is being manipulated at a time, the toy label is replaced with a quantifier indicating the number of toys involved in the action. To reduce the combinatorial factors, the quantifiers are limited to three choices of "two", "many", and "all" (see Appendix H for the complete coding manual). This labeling scheme results in an upper bound of 192 classes (24 *actions* $\times$ 5 *toys* + 3 *quantifiers* $\times$ 24 *actions* = 192 *classes*). Some combinations of quantifiers and actions are unlikely, and the data revealed 114 classes empirically.

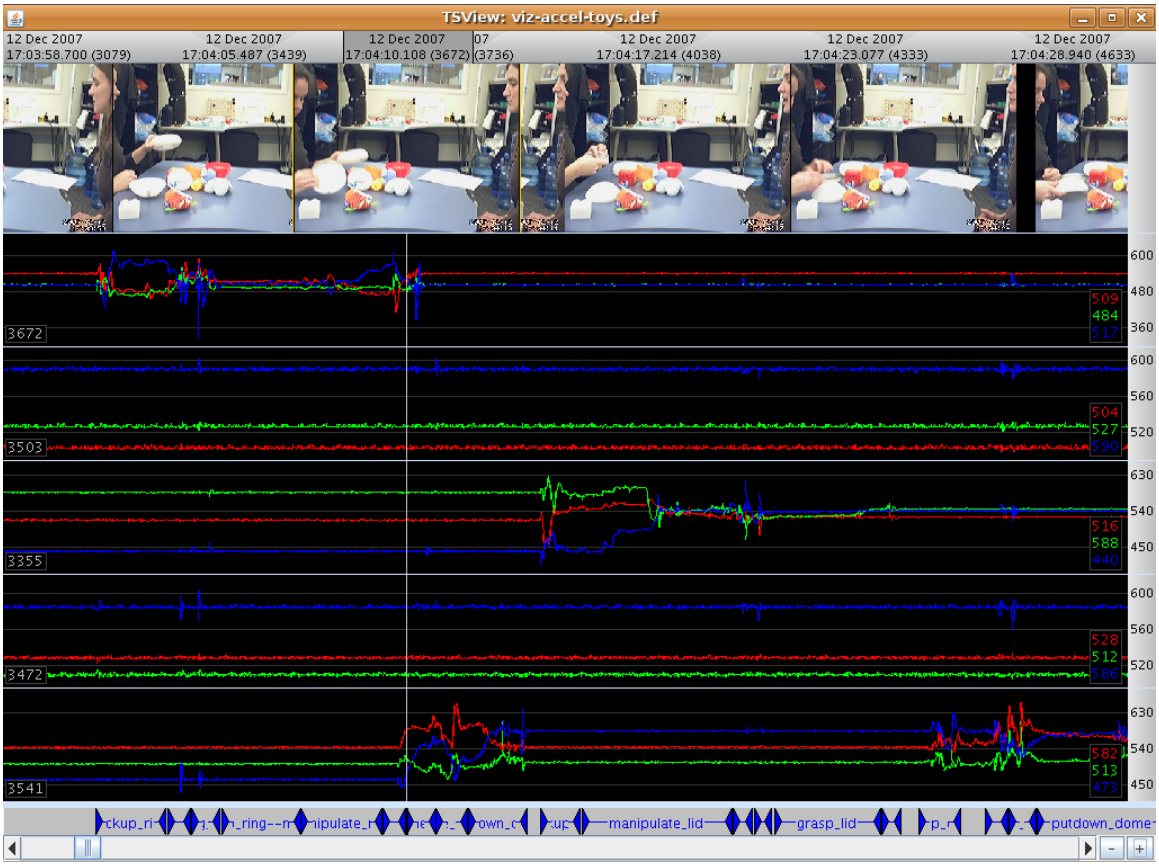


Figure 10: Screen capture from ground truth labeling software *TSview*

Table 6: Occurrence of 24 play primitives across all toys

Actions	Observed	Percent of Data	Duration in milliseconds		
			Minimum	Average	Maximum
bang	185	4.63 %	10	64	316
bump	53	1.33 %	4	41	412
drop	72	1.80 %	6	33	141
grasp	122	3.05 %	6	122	913
join	81	2.03 %	10	95	413
knockdown	2	0.05 %	51	123	195
manipulate	271	6.78 %	13	314	1631
move	254	6.35 %	12	65	416
pickup	755	18.88%	5	38	163
pour	26	0.65 %	102	189	344
push	332	8.30 %	7	53	289
putdown	646	16.15%	6	47	231
relate	49	1.23 %	49	244	890
release	33	0.83 %	4	28	256
reverb	163	4.08 %	2	108	584
roll	72	1.80 %	9	53	121
rub	7	0.18 %	115	223	329
separate	77	1.93 %	6	59	142
shake	129	3.23 %	15	91	418
spin	142	3.55 %	4	40	255
spinning	51	1.28 %	8	77	285
stack	228	5.70 %	8	63	404
takeout	82	2.05 %	21	67	376
unstack	167	4.18 %	5	41	146
Total	3999				

5.5 Features, Algorithms, and Analysis

Several steps are needed to prepare the raw accelerometer readings for analysis. First, although all of the accelerometers record data at the same frequency, the samples are not synchronized. Therefore each sensor stream is sampled at an even 50 Hz to estimate the instantaneous reading of each sensor at identical fixed intervals. Next, a one second window is slid along the 15D synchronized time series at $\frac{1}{3}$ second intervals. For each window, 315 features are computed including mean, variance, RMS, energy in various frequency bands, and differential descriptors for each dimension. Aggregate features are also computed based on each three-axes accelerometer including the mean, variance, and RMS of the magnitude of the sensor reading in three–dimensional–space and based on the angle of the vector to

the x-axis. This computation of aggregate features transforms a difficult temporal pattern recognition problem into a simpler spatial classification.

The models for recognition are trained using the iterative ADAboost framework where each iteration selects the best single feature and one–dimensional weak binary classifier for discrimination[60]. For our data set of 12 trials, 114 user–independent models were trained each using 30 rounds of boosting and leave–one–out 12–fold cross validation. The resulting ensemble binary classifiers consisted of both dual region decision stumps and tri–region Gaussian decision boundaries. As mentioned in Appendix D.4, there are several methods to combine multiple binary classifiers into a single multi–class classifier. Based on the results of previous work and empirical results, the one–vs–all multi–classificatoin scheme was used and the most probable model was selected via the probability summation method, *psum* (as described in Appendix D.4) [49]. During the classification process, a one second sliding window is again passed over the data. Thus, every $\frac{1}{3}$ second of data receives three classifications as it contributes to three distinct windows. A single classification for each $\frac{1}{3}$ second is derived by the *probsum* method (described in Appendix D.5) where the probability of each class given the window is calculated for any given one second window and the class with the largest probability sum is selected (see Figure 11).

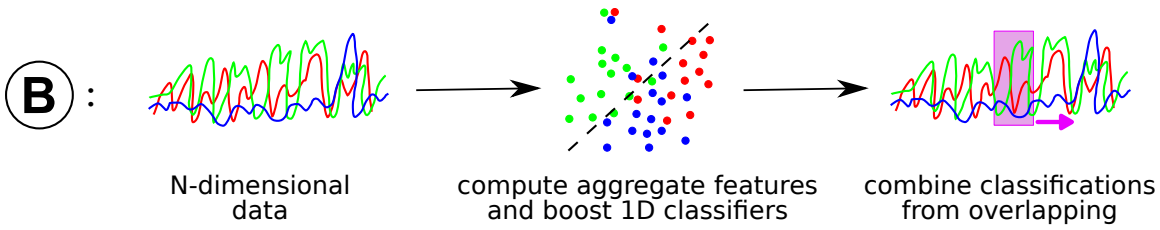


Figure 11: Summary of algorithm and parameters

5.6 Results

Several experiments were run involving different combinations of user–independent and user–dependent models as well as variations of toy–dependent and toy–independent models. Table 7 illustrates the results of user–independent, toy–dependent models. The event based matching criteria require that a continuous 10% of the event, at minimum, is labeled to be

Table 7: Average results of user–independent models for all toys and all actions

Evaluation	Accuracy	Units	
		Errors	Total
Events	25.3%	3856	5161
Frames	41.7%	6891	11823

classified as a correct instance, while the frame based evaluation accounts for correspondence of discrete, aligned chunks of label and ground truth. While it may seem counter–intuitive for the frame–level accuracy to be higher than the event–level accuracy, it is the result of both a combination of the event matching criteria and substitution–fragmentation type errors (see Appendix B for more details).

Table 8 shows the accuracy of the user–dependent, toy–dependent models for identifying play activities. As with the previous experiment, the frame–level analysis yielded higher accuracy rates than event level analysis. As one might expect, the average performance of the user–dependent, toy–dependent models is higher than the average performance of the user–independent, toy–dependent models.

Table 8: Results of user–dependent models for all toys and all actions

Participant	Evaluation	Accuracy	Units	
			Errors	Total
S1	Events	17.3%	615	744
	Frames	47.4%	687	1306
S2	Events	25.9%	665	898
	Frames	49.7%	1052	2090
S3	Events	25.4%	1185	1588
	Frames	44.6%	2037	3677
S4	Events	55.8%	199	450
	Frames	79.0%	233	1109
S5	Events	25.2%	747	998
	Frames	48.0%	1488	2864
Average		**29.9%**		
		53.7%		

In addition to evaluating the performance of the model with respect to the participant, models were also evaluated with respect to the toy. Table 9 shows the results of experiments

evaluating toy–independent and toy–dependent models. The first pair of experiments assesses the ability of the system to generalize models across the 24 primitive actions listed in Table 6. The toy dependent models, on average, showed a 6.6% absolute performance increase when evaluated by events and a 11.6% increase when evaluated by frames.

Table 9: Results of various user–independent model experiments that vary toy–independent and toy–dependent parameters

Experiment	**Evaluation**	**Accuracy**	**Units**		**Totals**		
			Errors	**Total**	**Toys**	**Actions**	**Models**
TD Actions	Events	25.3%	3856	5161	5	24	114
	Frames	41.7%	6891	11823	5	24	114
TI Actions	Events	18.7%	4212	5180	5	24	24
	Frames	30.1%	8259	11822	5	24	24

In addition to model experiments, I conducted an initial experiment using a naïve technique for classifying the presence or absences of relational play using the previous pilot data. The results in Table 10 have been obtained by computing the variance and pairwise Pearson's linear correlation coefficients over 333 millisecond sliding windows.

Table 10: Results of initial naïve binary classification

Ground Truth	Recall		Recognized		
	Hits	Percent	Total	Insertions	Precision
4	4	100.0%	36	32	11.11%
4	4	100.0%	67	63	5.97%
3	3	100.0%	40	37	7.50%
1	1	100.0%	102	101	0.98%
4	4	100.0%	57	53	7.01%
5	5	100.0%	101	96	4.95%
8	7	87.5%	117	110	5.98%
8	8	100.0%	98	90	8.16%
2	2	100.0%	98	96	2.04%
6	6	100.0%	77	71	7.79%
1	1	100.0%	128	127	0.78%
3	3	100.0%	141	138	2.12%
Average		98.96%			5.37%

The average recall rate is 98.96% with an average precision of 5.37%. The data used in

this experiment contained instances of both relational and non relational play, attributable to the high insertion rate and low precision. Several of the insertion errors were caused by two toys being manipulated independently, yet not in relation to each other. This naïve approach can detect the presence of motion in a toy and loosely correlate it to motion in other toys. However, it lacks the discriminative power to properly differentiate between the various types of motion required to identify relational play. Techniques that use naïve filters and models to recognize motion should see a reduction in number of insertion errors. These methods should also have the power to distinguish a larger variety of object play.

5.7 Discussion

The experiments involving the user–independent, toy–dependent models must distinguish between 114 classes. Selecting one class of the 114 classes at random would produce an accuracy of approximately 7.1% (assuming a uniform distribution of examples). Thus, the user–independent models, while demonstrating a seemingly low accuracy, improve recognition when compared to a random selection. The low accuracy of the models can be attributed to several factors. First, even though the NULL class accounted for a small fraction of the total data, the occurrence of any one single activity is extremely sparse. Longer activities, such as basic manipulation, had a higher accuracy per class than shorter activities, such as picking–up or putting–down a toy.

From Table 6 we can see that, on average, 70% of the activities have durations that are less than 100 milliseconds. The sliding window used in all of the above experiments was 1000 milliseconds long with an overlap of 333 milliseconds. As evidenced by the recognition rates, this window size may be too large to capture the nuances of the shorter play activities and may account for the decreased recognition accuracies. Furthermore, the sparsity of this data set may have proved insufficient to build proper models for recognition.

The free–form nature of the collected data does not ensure that there is an equal distribution of examples per play session. Hence, when training models, it is not guaranteed that each model gets an equal number of examples nor is it guaranteed that each session has all examples of the play behavior. When performing the leave–one–out 12–fold cross

validation, if the test set has a high concentration of play activities that is absent from the other trials, it both weakens the models for recognition and then presents a test set that is unrepresentative of the training data provided. Both factors can significantly reduce the accuracy of the models. Along these lines, the degrees of freedom per toy per activity may require many more rounds of boosting to allow adequate models to be constructed.

The performance of the models in the user–dependent case provides some evidence that recognition is possible to filter data for later review by an individual. For these sorts of visual inspection tools it is important to keep false–positives and false–negatives low (see Section 3.3). Humans may be able to quickly dismiss a few false–positives as irrelevant, but a significant number of them would more than likely detract from the system usability. Determining this exact number of errors and the impact of these errors will be the focus of the study proposed in Chapter 7.

5.8 Implications for Future Studies

The algorithm used in this pilot is well suited to on–body sensor systems [49, 44]. In such systems, the degrees–of–freedom of the sensors are limited to the kinematics of the body part to which they are attached. As this study has shown, the pilot method does not achieve as high of a level of accuracy (under similar training parameters) when applied to the unconstrained sensors within the augmented toys. Several questions arise as a result of the aforementioned experiments and will form the basis of studies discussed in the remainder of this dissertation:

1. Can recognition rates be improved by parameter modifications to the boosting method described in this chapter or are other methods more well suited towards this recognition task?

2. What features are best for describing object play? Should the features be based on single sensor streams or combinations of sensor streams?

3. Is it necessary to constrain the toys' freedom of movement to increase recognition rates or is increasing the number of training samples sufficient to improve

recognition?

4. What error rate do users find acceptable when using systems to support retrospective analysis?

The first three questions (Question 1 – Question 3) will be addressed by the study in Chapter 6. This study builds on the methods described in this chapter and will compare the recognition capabilities of this pilot algorithm to other common recognition algorithms, such as HMMs and SVMs, to answer the questions above. The final question (Question 4) will be addressed by the study presented in Chapter 7. This user–study has participants identify object play behaviors presented through an intelligent data visualization interface and quantifies the impact that various levels of recognition support have on this retrospective review task. Recent research has shown that users can tolerate accuracies as low as 60% when using gesture–based recognition systems [32], but it is unclear if retrospective review tasks will tolerate more or less errors.

CHAPTER VI

AUTOMATIC DETECTION OF OBJECT PLAY BEHAVIORS

The free–form nature of the data collected in the study presented in Chapter 5 is both a strength and a weakness of the pilot data set. Its unscripted nature and its combined use of augmented and non–augmented toys provides a more realistic data set. However, it also makes it difficult to characterize the performance of recognition algorithms over this data set due to the wide play variations found within and between participants. In this chapter I describe another adult object play data set that I collected which is more structured and involves play using only augmented toys. This chapter will not only describe the data set, but it also details how this data set is used to explore various feature spaces and recognition methods. The end goal of this exploration is to maximize the ability of the *Child'sPlay* system to recognize various types of object play as well as generalize to use on children's play data.

This chapter begins with the research questions addressed by this study and my hypothesis (Section 6.1). Next I provide a description of the adult and child data sets that are collected (Section 6.3) along with various feature spaces that are used to represent the data. The chapter concludes with presentation of recognition experiments and a discussion of the results(Section 6.5).

6.1 Research Questions and Hypothesis

The goal of collecting a larger adult play data set is to explore various feature spaces and recognition methods to help improve the recognition rates of certain object play behaviors as well as help support the overall identification of differing levels of play sophistication. In particular, this study is designed to address Research Question 2 and the following subquestions:

1. Can recognition rates be improved by using only augmented toys to collect object

play data versus a combination of augmented and off–the–shelf toys?

2. Which features spaces are best for representing object play behaviors.

3. How do effective retrieval rates of specific object play activities compare when adult object play is modeled using boosted ensemble classifies, hidden Markov models, and support vector machines.

I hypothesize that data from adults playing with augmented toys can be modeled using a combination of statistical methods, and the resulting models can be applied to data of children playing with the same toys. The recognition results on the children's data will have higher accuracy than the pilot approach discussed in Chapter 5 and allow for the identification of differing levels of play sophistication.

6.2 Data Collection Method

This section will describe the methods used to collect play data sets from both adult and child participants. The method used to collect the adult data set for these recognition experiments will be similar in many ways to the method used to collect the pilot data set in Section 5.3. However, there will be some important differences. First, the toys used during play will be restricted to augmented toys rather than the mix of augmented and off–the–shelf toys used in the pilot. Second, the play scenarios and promptings were scripted prior to the play sessions to help ensure that an equal and consistent number of play behaviors appear across sessions and participants. Also, the use of consistent prompts helped to create a data set that includes more frequent examples of higher levels of play involving multiple augmented toys, in contrast to the pilot data set. Third, adult participants will conduct play sessions while seated on a floor (rather than at a table) within a laboratory play space that has been designed to collect play data from infants and toddlers[1]. Adults played in the same play space as the toddlers to be more consistent with the toddler data set that was concurrently collected (see Section 6.2.2).

[1]This play space is the *Child Studies Lab* of the Health Systems Institute located on the campus of the Georgia Institute of Technology.

6.2.1 Data Collection from Adults

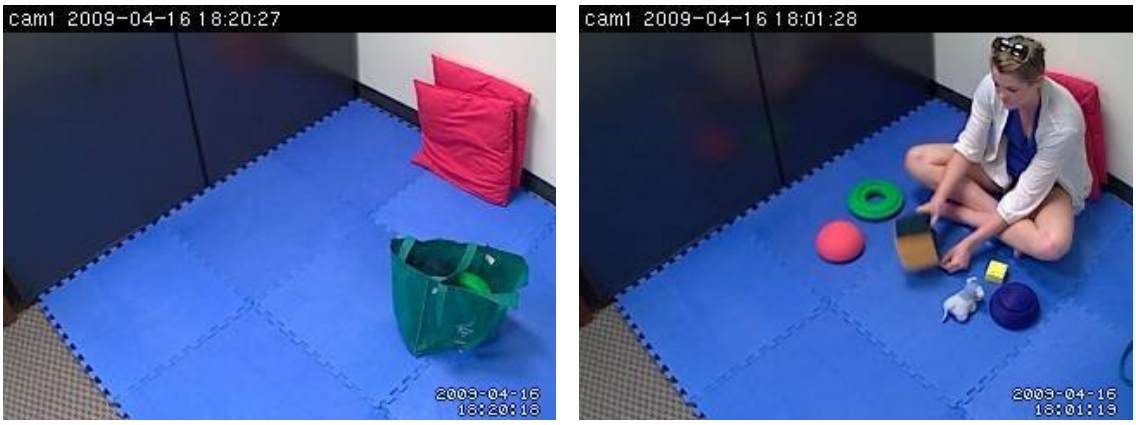


Figure 12: A view of the *Child Studies Lab* as shown from one of the overhead cameras. Left: the play space prior to the start of a session. Right: the play space while adult play data is being collected

Ten able–bodied adults were recruited, four females and six males. Each subject participated in four play sessions, with each session occurring on a different day. Data was collected from April 2, 2009 – April 30, 2009 with participants' play sessions ranging from 25 – 35 minutes. At the start of the session, each participant is asked to sit in the child play space and is then presented with an opaque bag containing the seven augmented toys (see Figure 12). Participants were asked to shake the bag twice (to provide a means of verifying that the data is synchronized during post processing) and then were asked to retrieve toys from the bag. Some participants chose to dump the contents of the bag onto the floor while others removed toys in an orderly fashion from the bag. Once the toys were on the floor, the participants were instructed to perform a series of play tasks. At the end of each session, toys were placed back into the bag, and the bag was shaken once again for data synchronization purposes. Afterwards, the participants were again asked to remove the toys from the bag and were instructed to "play with the toys however they liked." This free play session typically lasted 3 minutes, though there was no set time limit. When the adults were done playing, the toys were once again placed in the bag and shaken for the purposes of data synchronization. It should be noted that although adult free play sessions were collected, recognition results will only be reported for the scripted portion of the play session to ensure uniformity of samples between participants.

Table 11: Play procedure data collection sheet for adult participants

	Exploration of Objects in Play						Relational Use of Objects in Play		Functional Use of Objects				
	Indiscriminate Actions			Manipulations of Single Objects			Takes Combinations Apart	Presentation and General Combinations	Semi Object Directed	Semi Self Directed	Playing with Puppy in Motion		
Toys	**explore**	**shake**	**bang**	**rock**	**spin**	**roll**	**take apart**	**assemble**			**puppy nested-ground**	**puppy ground motion**	**puppy nested-air**
Ring	Find the yellow dot on the ring	Does the ring have a rattle?	use the ring to hammer in an imaginary nail	*hold puppy by the head and thrash everything*	can you make the ring spin around like a record?	can you roll the ring	build the flying saucer; take apart ring and red dome	have trouble but put the red dome on the green ring	build the flying saucer and have it fly	put on the braclet	sit the puppy in the ring	have the puppy run towards the ring and jump over it	sit the puppy in the ring and fly him through the air
Dome (red)	Find the green dot on the red dome	Does the red dome have a rattle?	Use the red dome to hammer in an imaginary nail	can you make the red dome seesaw?	can you make the red dome spin?	can you roll the red dome like a tire?	build the ball; take apart the ball	have trouble but put the red dome on the blue dome	build the ball and roll it around	drink from the red cup	sit the puppy in the red dome, rock the dome	have the puppy run towards the red dome and jump over it	sit the puppy in the red dome and fly him through the air
Dome (blue)	Find the green dot on the blue dome	Does the blue dome have a rattle?	Use the blue dome to hammer in an imaginary nail	can you make the blue dome seesaw?	can you make the blue dome spin?	can you roll the blue dome?	build a tower with the domes; take apart stacked domes	have trouble but form a ball with the domes	build the ball and toss it inbetween your hands	drink from the blue cup	sit the puppy in the blue dome, rock the dome	have puppy run towards the blue dome and jump over it	sit the puppy in the blue dome and fly him through the air
Lego (yellow)	Find the blue dot on the lego	Does the lego have a rattle?	use the lego to hammer in an imaginary nail	*hold puppy by the head and thrash everything*	*spike the puppy face down like a football*	can you roll the lego	put the legos together; take apart legos	have trouble but put the legos together	0	put the legos together	have puppy push the lego like a bull dozer	have puppy run towards the yellow lego and jump over it	sit the lego in the blue dome and fly him through the air
Lego (grey)	Find the blue dot on the other lego	Does the other lego have a rattle?	use the other lego to hammer in an imaginary nail	*hold puppy by the head and thrash everything*	0	can you roll the other lego	put the legos together; take apart legos	have trouble but put the legos together	0	0	have puppy push the other lego like a bull dozer	have the puppy run towards the other lego and jump over it	sit the other lego in the blue dome and fly him through the air
Puppy	Find the red dot on the puppy	Does the puppy have a rattle?	use the puppy to hammer an imaginary nail	*hold puppy by the head and thrash everything*	0	can you roll the puppy	0	*hammer the puppy with the ring*	have puppy run around and bark	have the puppy run to you and bark	stack the legos on their side, have puppy run to it and knock them down	have the puppy run around and bark	have the puppy run to the lego and ram it
Block	Find the white dot on the block	Does the block have a rattle?	use the block to hammer an imaginary nail	*hold puppy by the head and thrash everything in site*	0	can you roll the dice	0	make a hamburger with the red dome, the green ring, and the blue dome	roll the dice	roll the dice	have puppy push the block like a bull dozer	have the puppy run towards the block and jump over it	have the puppy run to the other lego and ram it

As with the pilot data set, the instructions provided during each sessions were designed to elicit certain play behaviors without directly asking for specific activities to be performed (when possible). Each session consisted of 83 scripted play prompts. Prompts were selected at random from the protocol sheet and marked once performed to prevent accidentally reusing the prompt. Table 11 provides the protocol sheet used when collecting adult data (a larger version is provided in Appendix G). The column headings specify the basic play activity (organized loosely by level), and the row headings indicate the primary toy used in the interaction. Each cell of the table indicates the prompt to be provided to the participant. For example, the prompt "use the red dome to hammer in an imaginary nail" is used to elicit examples of banging with the red plastic dome. A more subtle example is the prompt designed to elicit exploratory examinations of toys. While children have a tendency to explore unfamiliar objects, adults sometimes need more coaxing. Rather than simply asking the adults to examine the yellow Lego^{TM} Quatro, the prompt asks the adult to "find the blue dot on the lego." However, some of the toys have dot stickers on them while others do not (often changing from session to session). The variability in dot location usually causes the adult to carefully examine the toy during each session.

The last three columns of the protocol sheet are under the general heading "Playing with Puppy in Motion." These actions were included after several meetings with a developmental psychologist who helped review and revise this play protocol. Activities in this category are higher-level, developmentally-relevant activities in which children are likely to engage given the set of augmented toys [4]. Typically developing children like to place toys inside of other toys (loosely referred to nesting on the protocol sheet) and the abstract nature of the augmented toys lend themselves to these types of behaviors. In particular, it is important to detect motions with the plush puppy rattle as it has a social face and common functional/imaginary uses for children. The "Playing with Puppy in Motion" category is designed to collect examples of the plush puppy rattle being nested inside objects in contact with the ground; being made to run around as well as interact with other toys; and being nested inside objects that fly through the air (see Figure 14). While these puppy motions are more sophisticated and not likely to be demonstrated by the age group targeted by the *Child'sPlay* system, they are included in the data set to help promote future pattern recognition research.

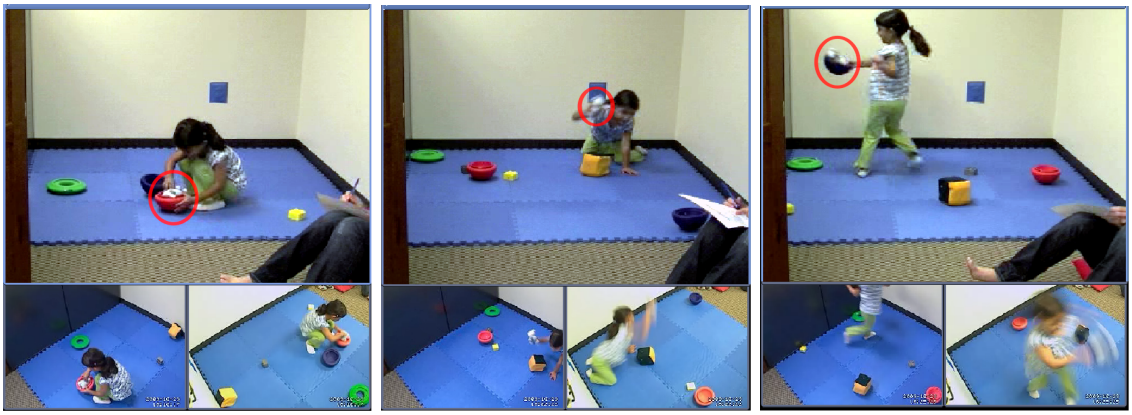


Figure 13: Three examples of the plush puppy rattle toy being used in play as seen by three overhead cameras. The puppy has been circled in red. Left: puppy is nested in the red dome; Middle: the puppy is running and jumping over another toy; Right: the puppy is flying while seated inside the blue dome.

In addition to appropriate play, it is also important to have examples of atypical play. In particular, it is of value to collect data when social toys, such as the puppy, are misused [4]. An example of misuse would be when the social properties of the puppy are ignored. For

example, grabbing the puppy by the face, dropping the puppy on its face, or purposefully throwing the puppy down face first. Another sign of inappropriate play is when the puppy is used in an indiscriminate manner and thrashed against other toys. Items in red italics on the play protocol sheet indicate negative play behaviors[2]. In particular, participants are instructed to hammer the puppy in the head with the ring; spike the puppy down on its face; and grab the puppy by the head while thrashing wildly it into other toys.

It should also be mentioned that while the play protocol sheet is designed to gather a data set with an equal number of activity per toy per participant, it does not strictly guarantee uniformity across the data set. For example, adults often fidgeted with toys in between prompts or performed the prompt more than once. For example, when rolling one of the domes across the floor, participants often rolled it more than once to get a "good roll" or toss the lego from hand–to–hand as they put it down. The protocol does, however, help make the data set more uniform than the loose prompting method used to collect the pilot data set.

6.2.2 Data Collection from Children

In addition to the adult data set I have collected three different object play data sets with a total of thirteen child participants. These data sets are the *Rapid ABC play* data set, the *multi-visit* data set, and the *HSI Parent Infant Social Games Video Library play* data set. Each of these data sets were collected in the *Child Studies Lab* Play Space.

6.2.2.1 Rapid ABC play data

Play data from the eight toddlers was collected from April 3, 2009 – December 21, 2009. These children ranged in age from 15 – 36 months of age. One of the eight has been classified as "at–risk" for a future diagnosis of autism spectrum disorder. Play sessions ranged from five to ten minutes.

The eight children, 6 girls and 2 boys, were dual recruited in association with a pilot test of the Rapid Attention Back and Forth Communication (Rapid ABC) — an assessment

[2]Negative play behaviors do not necessarily correspond to the overall activity column in which they appear.

designed to help identify children at risk for autism spectrum disorder. This autism screening assessment, targeted at children between ages 15 – 27 months of age is part of a joint collaboration between the Emory Autism Resource Center (EAC) and the Health Systems Institute (HSI) at Georgia Institute of Technology. When enrolling their child in the Rapid ABC pilot, parents can choose their assessment location to be at either HSI or EAC. Parents that enrolled at the HSI location were also given the option to enroll their child to be a participant in the *Child'sPlay* data set and have their child play with augmented toys. The Rapid ABC assessment at HSI is conducted in the Child Studies Lab, in the same location as the *Child'sPlay* play space. Children that were dual recruited first participated in the Rapid ABC assessment. The Rapid ABC lasts 3–5 minutes and consists of five socially oriented tasks. After completing the assessment, the child then plays with the augmented toys as their parent completes surveys and questionnaires administered by the Rapid ABC clinician. While playing, the child is always within line of sight of his parent and the Rapid ABC clinician. I was supervising the child directly while they were in the play space.

Prior to the arrival of the child, the toys are placed in an opaque bag and shaken three times for data synchronization. When the child enters the play space, the toys are in the bag, located at the center of the play space. For these play sessions there was no scripting, the child is merely encouraged to explore what is in the bag and encouraged to find a new toy when they become bored with the current toy with which they are playing. As the supervisor and primary play participant with the child during this time, I would often engage the child in social games to encourage play.

6.2.2.2 Multi-visit

Two children, a boy and a girl, were recruited independently from the Rapid ABC pilot study and played with the augmented toys over multiple visits. Data was collected from the boy, age 31 months, over four sessions, each lasting at least 20 minutes. There were five months between his first and second session (April 2009 – October 2009); his second and third sessions occurred during the same week in October; and his final session occurred a month later, in November 2009. During all sessions the child was accompanied by a parent,

and the parent served as the primary play partner for the child. The parents were asked to encourage their child to play with the augmented toys while avoiding direct contact with the toys themselves, if possible. The parents were asked to direct attention to toys by pointing or asking the child questions pertaining to characteristics of the toys. The sessions contain a variety of object play activities as well as social games involving use of the toys.

The girl, age 5, had data collected during four visits spanning two weeks in October 2009. The data from her play sessions was collected specifically for use in the retrospective review study discussed in Chapter 7. To help ensure consistency of play between the sessions, the girl participated in the adult play protocol. It should be noted that a younger child was originally recruited for this data set; however, the child was unable to follow the instructions in a consistent manner.

As with the adults, each session consisted of 83 scripted play prompts that were selected at random (without repetition) from the protocol sheet (see Table 11). The girl was the only person in the play space while collecting the data. Although the adult protocol was used, the length and nature of the play is very distinct from the adult subjects. Like the adults, the girl often fidgeted with the toys between prompts. However, the girl often used her whole body to interact with the toys during these fidgets. For example, she would often sit on the plastic dome toys and rock or slide across the floor while seated in them. Furthermore, she would often run around the play space, hopping and jumping between prompts. These extraneous motions were often registered by the sensors inside the toys even though they were not directly used.

The nature of the play itself also differed from the adults. For example, when asked to wear the green ring as a bracelet, she would often place the ring on her arm and then immediately remove it and place it on her leg, claiming that she liked it better in that position. As expected, there were also larger variations in the way she performed her play activities. For example, when asked to make the puppy run around, sometimes the puppy would run across the ground similar to the adults. However, sometimes she jumped across the floor holding the puppy such that he hit the ground when she landed each time. When she was asked to fly the puppy around in the domes, she would sometimes spin rapidly

holding the puppy and dome outward. Other times, she would raise the puppy and dome up in the air vertically and hold them stationary in the air.

It should be noted that the girl was recruited because of her age. Younger children (30 months) were recruited, however, these children were too young to be able to complete half a session of the adult protocol. Even at 5 years of age, the girl showed signs of frustration in some of the latter sessions with the repetitive nature of tasks.

6.2.2.3 HSI Parent Infant Social Games Video Library

This data set was collected in collaboration with Wang *et al.* (see Section 2.2.3) [68]. In this data set three children, ages 2, 4, and 4, each played a series of social games with their respective parents. Both the augmented toys and several off–the–shelf toys were played with during these social games. This data set consists of 85 minutes of contiguous play data. Analysis of this data set is beyond the scope of this dissertation. This data was collected to support future research involving a combination of vision based techniques and augmented toys to help automatically characterize a wide variety of play activities. This fusion of techniques will be discussed in more detail in Chapter 8.

6.3 Description of the Data Collected

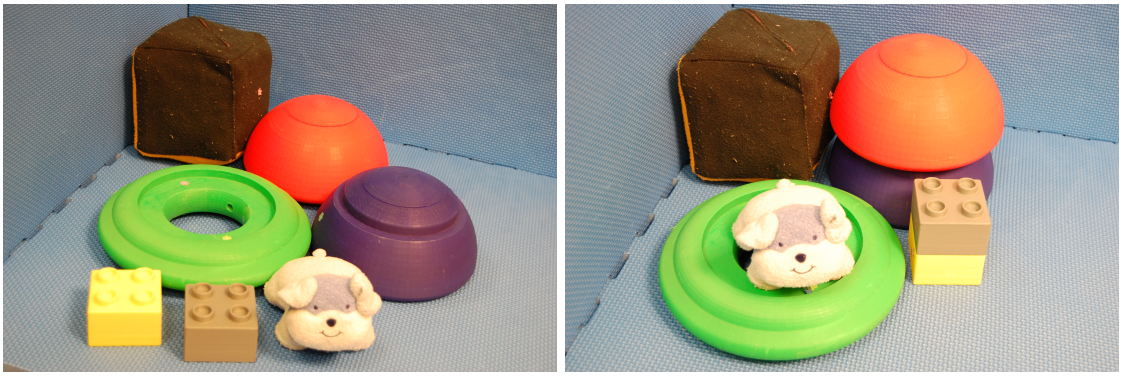


Figure 14: The augmented toys of the *Child'sPlay* system. Left: individual toys. Right: toys assembled.

The seven augmented toys (two plastic domes, one plush cube, one plush puppy rattle, one plastic ring, and two Lego^{TM} Quatro) bricks described in Section 4.3 were used to collect data at each of the play sessions. Each augmented toy contains a BlueSense sensor which

has an integrated 3–axis accelerometer, dual–channel capacitance sensor, and single channel sound processing capabilities. However, only the accelerometer data is recorded during the play sessions. Therefore, each toy produces three–dimensional data, and a session with seven toys will produce 21–dimensional data. While the BlueSense sensors are capable of producing data rates as high as 160 samples per second, the sensors were reprogrammed to sample at 40 Hz. The reduction in data transfer helps prevent the data recording system from dropping packets. Empirically, frequencies associated with object play did not exceed 17 Hz. Therefore, a 40 hz sampling rate is appropriate, according to the Nyquist–Shannon sampling theorem [61].

In addition to collecting BlueSense data, motion jpegs from three overhead network cameras, high–definition audio–video from a frontal view, and environmental audio from a single microphone (positioned above the floor) is collected during each play session. As with the pilot data set described Section 5.4, both the BlueSense data and audio–video data are collected on the same machine to simplify post–process data synchronization. The high–definition video, however, was captured on an independent camcorder, and the start point of the video must be manually synchronized with the start of the other video feeds.

Labeling of the adult data set began on May 09, 2009. Six students were recruited independently to exhaustively label portions of both data sets using the *PlayView* software (see Figure 15 — The *PlayView* interface will be described in more detail in Chapter 7). These students were trained for approximately 1.5 hours to identify the object play activities of interest and for 20 minutes to use the *PlayView* software. During training, each person was provided with a coding manual (see Appendix G) indicating the 38 object play activities to identify within the data (see Table 12).

The data coders were asked to log their progress to help tabulate how much time they spent labeling each play session. However, none of the coders consistently logged their progress, stating that it was too much of a burden while labeling the data. The data coders self–reported (verbally) that they took 3–5 hours, on average, to label a 30 minute adult play session. They reported that the children's sessions take longer. The data coders also noted that the visualizations of the acceleration data sensor streams were very helpful in

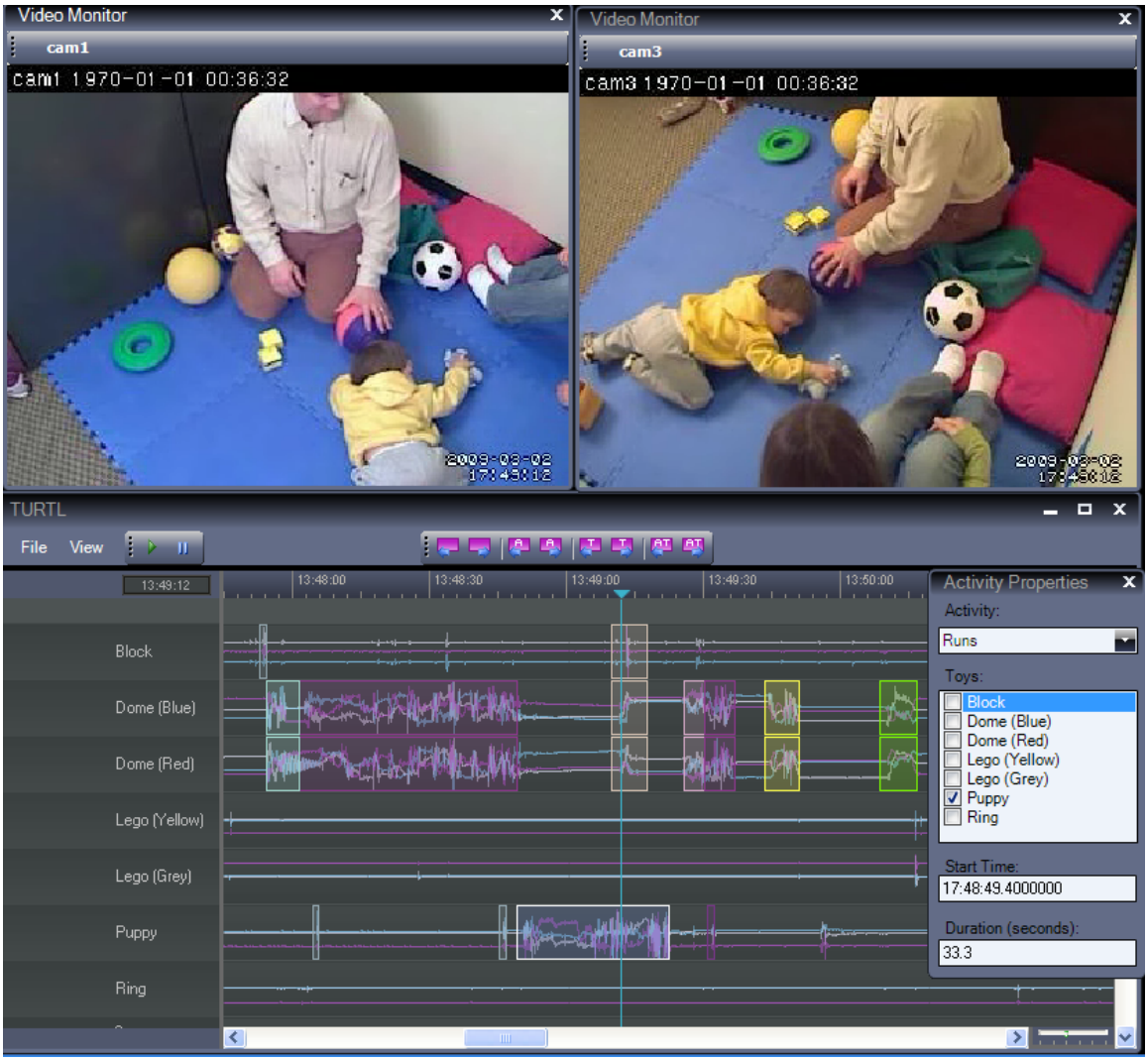


Figure 15: Screen capture of the *PlayView* interface used to label the data sets

the labeling process.[3]

In contrast with how the pilot data set was labeled, each sensor stream is labeled on a per–stream basis, rather than using a single label to represent all seven streams for any given instance in time. As a consequence, the data set has overlapping labels. Any of the 38 activities can be applied to a majority of the seven toys as well as combinations of assembled and nested toys. Empirically, there were 311 toy–dependent activities identified within the adult data set and 269 toy–dependent identified within the child data set.

Cohen's kappa coefficient was computed to determine the inter–rater agreement between

[3]These visualizations may have also helped increase inter–rater agreement.

two data coders [15]. The two coders were given identical portions of play data that accounted for approximately twenty percent of the total adult data set. The labels provided by the two data coders exhibited agreement for 311 toy–dependent activities with an average Cohen's kappa of $\kappa = 0.61$. According to the scale presented by Landis and Koch for interpreting κ coefficient values, the experimental determination of $\kappa = 0.61$ indicates moderate to substantial agreement amongst the two raters[41]. Many of the object play activities, such as shaking the plush puppy rattle, had substantial to near perfect agreement. Non–play activities, such as fidget or bump, often had fair to no agreement at all which lowered the overall average κ. For the purposes of the *Child'sPlay* system it is acceptable to have confusion among the non–play activities as the *Child'sPlay* system is currently focused on distinguishing between different types of object play activities and does not need to differentiate between various types of non–play.

When conducting the recognition experiments, non–play activities, such as bumping toys, fidgeting with toys, and toy reverberations, are grouped into the *NONE* class. In addition, slight variations between coders are accounted for by filtering out activities that have less than five total examples across the entire data set. As such, the adult data set consists of 96 toy–dependent classes and the child data set consists of 160 toy–dependent classes. These classes as well as, their average duration and the percentage of the data set for which they account are listed in Table 12 and Table 13.

6.4 Feature Selection and Data Modeling

Crucial to the success of any recognition system is appropriate feature selection and model selection. In selecting an appropriate representation, it is important to understand what behaviors are being modeled (as well as why) and select methods that best represent these aspects. When characterizing early object play, play can have both periodic and aperiodic properties. For example, in the earlier stages of development, toys can be shaken, which involves a repetitive back–and–forth motion. Or toys can be explored, which can be as simple as single touch of the object, or, as complex as repeatedly rotating an object while visually inspecting it. The speed and precision of repetitive activities is likely to increase

Table 12: Occurrence of play primitives across all toys in 34 play sessions of the adult data set

Actions	Observed	Percent of Data	Duration in milliseconds		
			Minimum	Average	Maximum
non–play (NONE)	3255	29.35 %	35	196	3448
assemble	535	4.82 %	20	172	1159
bang	233	2.10 %	24	184	475
drinks	65	0.59 %	88	194	412
drop	19	0.17 %	15	68	204
explore	306	2.76 %	30	322	1166
falls	12	0.11 %	27	42	71
flies	215	1.94 %	53	282	754
hammers	34	0.31 %	85	207	722
jumps	217	1.96 %	5	42	154
knocks down	15	0.14 %	21	44	136
nest	282	2.54 %	16	73	271
pushes	160	1.44 %	11	224	833
rams	157	1.42 %	14	85	232
relate	252	2.27 %	9	207	1159
relocate	1605	14.47 %	7	49	246
rock	149	1.34 %	23	513	1970
roll	565	5.09 %	8	165	1252
runs	352	3.17 %	13	90	689
runs towards self	62	0.56 %	59	195	594
separate	541	4.88 %	10	74	369
shake	286	2.58 %	24	130	400
slide	1196	10.78 %	8	29	751
spikes	37	0.33 %	36	119	266
spin	216	1.95 %	17	342	1556
stacks	67	0.60 %	21	83	244
thrashes	168	1.51 %	51	292	816
toss	57	0.51 %	16	177	599
wears	34	0.31 %	132	235	573

Total Play Events	7837
Total Overall Events	11092
Bias towards Play	70.65%
Bias towards Non–Play	29.35%

over time as the child's motor skills and coordination develops. Aperiodic play motions, such as grasping an object, putting on a plastic bracelet, or taking an imaginary drink from a plastic cup also occur during play. Therefore, both the features selected from the data,

and the models used to represent play activities must be able to account for both periodic and aperiodic motions as well as accommodate temporal variations. Also, as discussed in Section 6.4.1, representations that can help distinguish play with single objects as well as play involving multiple objects is important to categorizing different levels of object play. This section will describe several different feature combinations and model representations that were explored for developing the play recognition portion of the *Child'sPlay* system.

6.4.1 Feature Selection

Before discussing the specifics of the feature selection process, I will briefly describe the pre–processing that occurs with the data. As with the pilot data, several steps are needed to prepare the raw accelerometer readings for analysis. First, each sensor stream is sampled at an even 35 Hz to estimate instantaneous readings at identical fixed intervals across all sensors. Second, a one and a half second sliding window is applied to the 21–dimensional synchronized time series data at half second intervals. Therefore, each window consists of 50 samples of data and overlaps with two neighboring windows by 33 samples. Several different features are then computed over these windows.

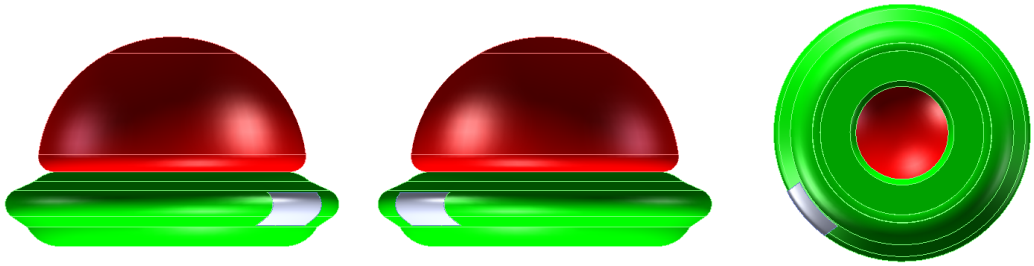


Figure 16: Illustration of the effect of rotation on sensor orientation. The location of the sensor within the ring toy is highlighted white to demonstrate the effects of rotation about Z–axis (Left and Middle) and to highlight the off–axis sensor placement within the plastic ring toy (Right).

Similar to the pilot data set, simple descriptive statistics are computed for each window. In particular, the mean, the second through fourth central moments (variance, skew, and kurtosis), and the change in variance are computed for each window. When all are used,

these statistics represent 105 elements of the final feature vector (21×5=105). The approximations of orientation that are included in the pilot studies, however, are not computed for this feature set. Except for the plush puppy rattle, the toys are abstract in shape and orientation information cannot be used to determine appropriate versus inappropriate use. Orientation approximations, such as rotation about the toy's central axis, often add more noise than information as the round toys can be grasped and used in any number of orientations. This problem is further exacerbated for the plastic ring toy, where the sensor is not aligned with the central axis of the toy. The large hole in the center of the plastic ring toy forces the sensor to be nested off–axis (see Figure 16). For this reason, several rotationally invariant features are added to the features used in the original pilot studies. Measures such as the entropy, power spectrum densities and correlative measures, are also computed for each dimension.

The power spectral density function is a rotationally invariant metric that can easily represent periodic play motion as a variation in power over a given range of frequencies (see Figure 17). The power spectral density features are computed per accelerometer axis using a 32–point fast Fourier transform. Only the first half of the frequency bins (1–9) are retained. Information for frequencies 10 – 17 are discarded as, empirically, object play motion seldom reaches frequencies in this range. Because our sensors sample at 35 Hz, 17 Hz is the highest frequency for which valid information can be retrieved according to the Nyquist–Shannon sampling theorem. However, even when only using the first 9 frequencies, computing power spectral density features for each axis of each toy would lead to an additional 189 features (7 toys × 9 frequencies × 3 axes = 189 features). To help reduce dimensionality, the power spectral density features for the three axes ($x_{psd_{t_n}}, y_{psd_{t_n}}, z_{psd_{t_n}}$) corresponding to the same toy, t_n, are combined into a single density measurement by computing the quadratic mean, $\mu_{psd_{rms}}$ (the root mean square – see Equation 1). Therefore, the power spectrum density for each toy is reduced from 27 features per toy to 9 features per toy resulting in 63 features total.

Figure 18 illustrates the similarity of the power spectral density features of different toys being shaken at different times. Power spectral densities may be a sufficient feature

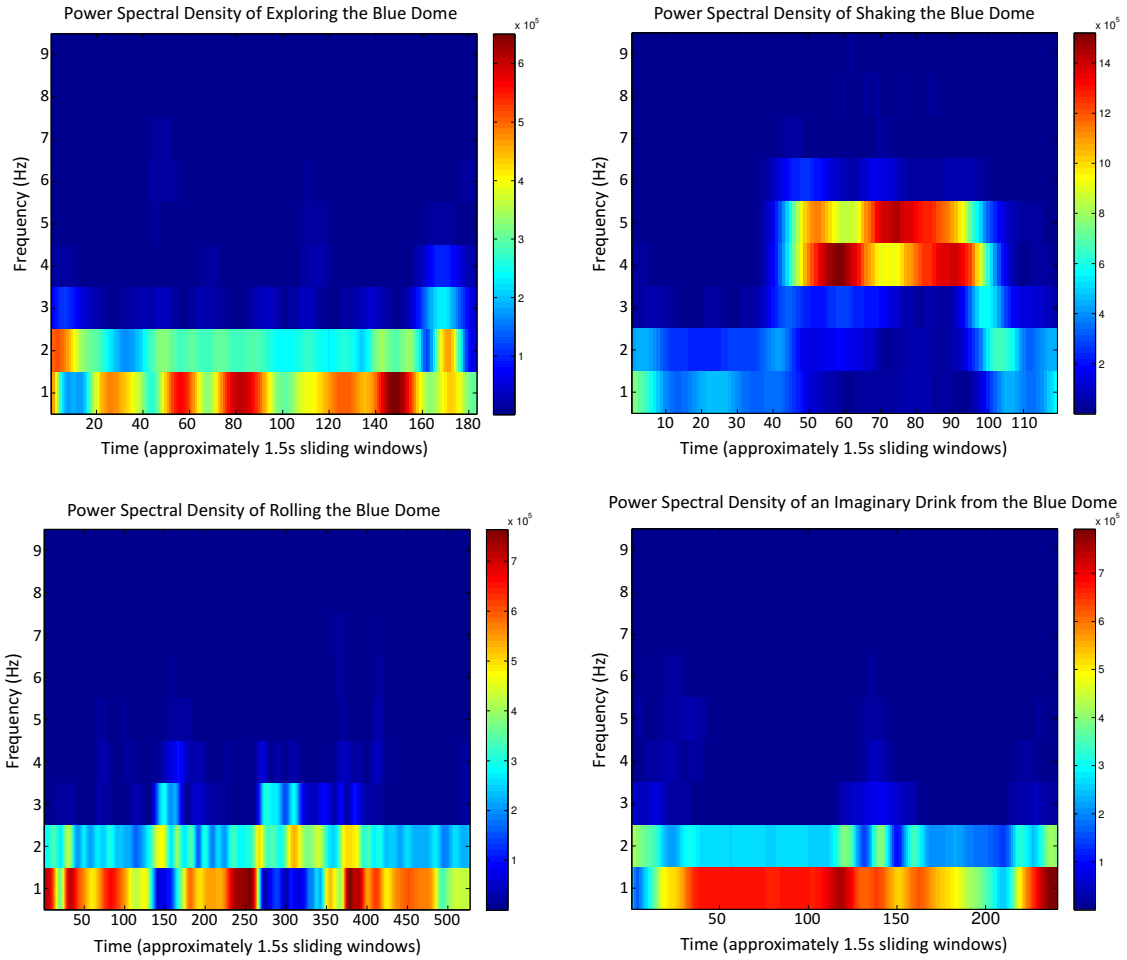


Figure 17: A comparison of the power spectral density computed over the blue plastic dome while being explored (top left), shaken (top right), rolled (bottom left), and used as an imaginary drinking cup (bottom right).

for representing activities independent of the toy involved in play. For example, shaking consistently has power variations in the 4 Hz to 5 Hz range, regardless of the toy being shaken. Likewise, Figure 17 shows the differences between various play activities performed using the same toy. In this collection of images, the periodicity of shaking is contrasted against other periodic play motions (such as rolling the dome and exploring the dome). The distinction between the periodic shaking and the aperiodic motion of imaginary drinking can be seen as well.

$$\mu_{psd_{rms}} = \sqrt{\frac{x_{psd_{t_n}}^2 + y_{psd_{t_n}}^2 + z_{psd_{t_n}}^2}{3}} \tag{1}$$

In addition to the power spectral densities, aggregate features are computed over pair–wise combinations of the sensor streams to help determine if toys are being manipulated in a similar fashion.

Features based on coherence, correlation, and cross–covariance are all useful metrics for determining the similarity between time–series data produced by accelerometers [43, 48, 5]. In particular, Pearson pairwise linear correlation coefficients are calculated between raw sensor readings for every axis, resulting in 189 features. Note that there are 210 possible combinations for seven sensors with 3 axes each, however, 21 combinations can effectively be ignored as they represent correlations between axes on the same sensor. Correlation coefficients are also computed between the power spectral densities (approximating the cross power spectral density) resulting in 252 additional features.

If all of the features described above are to be used simultaneously, each 1.5 second window will be represented by a feature vector consisting of 610 elements. Given that the adult data set alone consists of 87947 windows and 96 classes when recognizing toy dependent activities, memory and time requirements for algorithms for feature selection or model inference can quickly become computationally prohibitive. Therefore, many of the experiments discussed in Section 6.5 will involve subsets of the features described above.

6.4.2 Data Models

All of the features computed in Section 6.4.1 are calculated over sliding windows. The use of sliding windows to create aggregate features transforms a temporal pattern recognition problem into a simpler spatial classification and allows for a variety of supervised and unsupervised learning techniques to be explored. Recognition experiments were conducted using ensemble classifiers (Section 6.5.1), hidden Markov models (Section 6.5.2), and support vector machines (Section 6.5.3).

For comparative purposes with the pilot recognition experiments (described in Section 5.6), initial model exploration experiments are conducted using fourteen of the forty play sessions. Although the experiments are now matched for fourteen play sessions, it should be noted that these comparisons may be slightly misleading as the two data sets

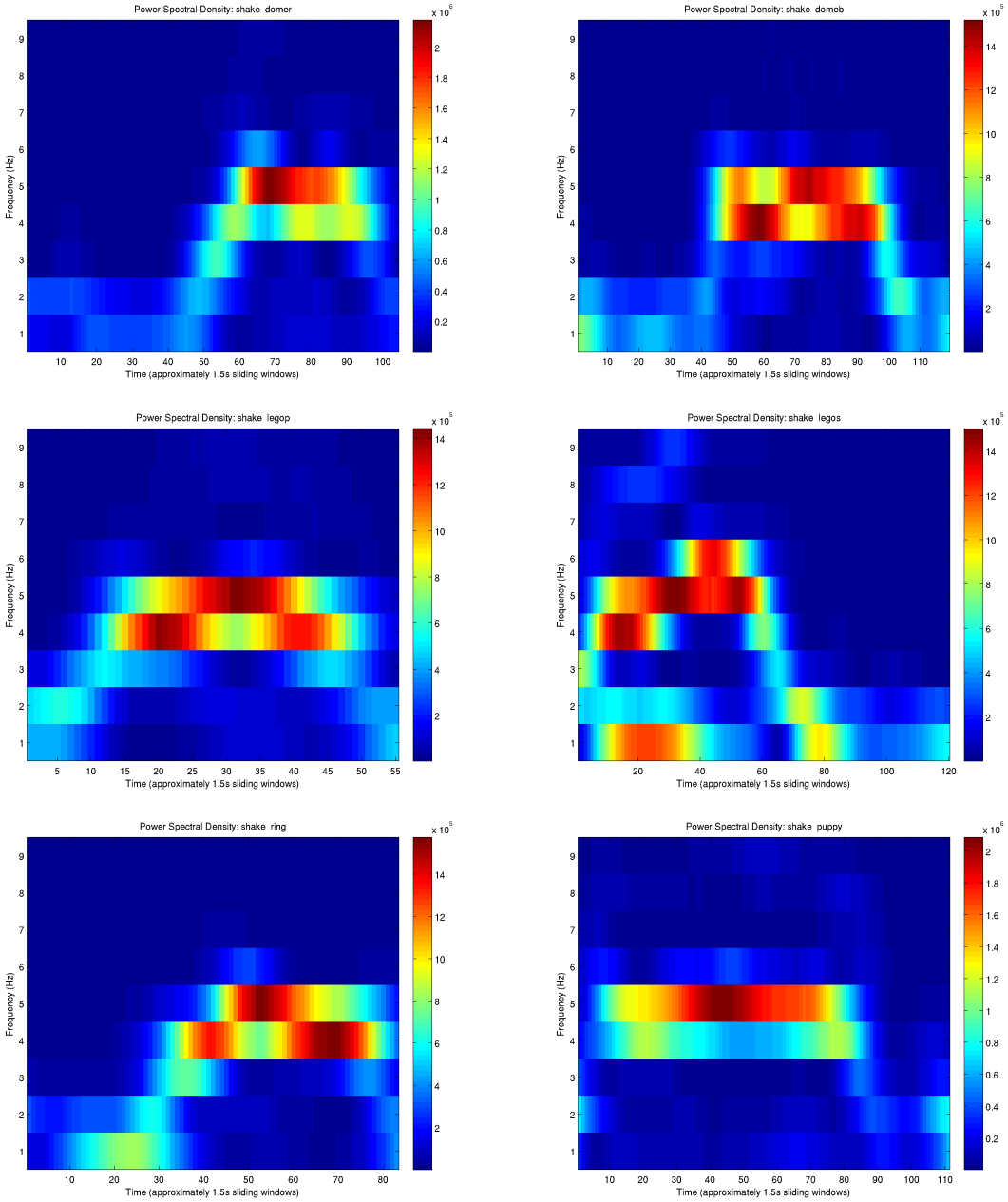


Figure 18: A comparison of the power spectral density computed over six different toys while being shaken: the blue plastic dome (top left), red plastic dome (top right), the yellow Lego^{TM} Quatro (middle left), the grey Lego^{TM} Quatro (middle right), the green plastic ring (bottom left), and the plush puppy rattle (bottom right).

were collected using different augmented toys and spanned a slightly different set of play activities (see Table 12 and Table 6). The model which performs the best, comparatively, will be applied to the children's data set. However, when conducting experiments to test the generalization of adult models to recognize children's play activities, all of the adult play sessions will be used as training examples instead of the fourteen session subset.

6.5 Results

This section includes results from the various experiments described in the previous section. First, the results from models constructed over a subset of data with size comparable to the pilot data set will be presented, followed by the presentation of experiments testing the ability of models trained on adult data (the adult models) to generalize to the children's play data. Due to the sparsity of object play events within the data, the accuracy metric is not always the best metric to measure system performance for correctly identifying object play activity. The standard accuracy metric assigns equal weight to true positives and true negatives (see Equation 2). As such, identifying a large number of non–play events correctly can result in a high accuracy even if relatively few object play events are identified correctly. To obtain a more comprehensive picture of system performance, all experimental results will be reported in terms of accuracy, true positive rate (recall), false positive rate, specificity, positive prediction value (precision), negative prediction value, false discovery rate, and the F_1 score. These metrics are defined and described with more detail in Appendix B.4. The F_1 score is the harmonic mean of precision and recall (see Equation 3) [58, 30]. The F_1 score measures the ability of the system to effectively retrieve specific play activities and more accurately reflects overall system performance as related to retrospective review tasks. A high F_1 score implies that the system is good at both detecting all events and avoiding false detections. Therefore, for the experiments presented in this chapter, the F_1 score is used as the metric with which to compare model performance overall.

$$Accuracy = \frac{((True\ Positives + True\ Negatives) - False\ Positives)}{(Positives + Negatives)} \tag{2}$$

$$F_\beta = (1 + \beta^2) \cdot \frac{Precision \cdot Recall}{\beta^2 \cdot Precision + Recall} \tag{3}$$

$$True\ Positive\ Rate\ (Recall) = \frac{True\ Positives}{Positives} \tag{4}$$

$$Positive\ Prediction\ Vale\ (Precision) = \frac{True\ Positives}{(True\ Positives + False\ Positives)} \tag{5}$$

6.5.1 Boosting One–Dimensional Classifiers

As a baseline, the boosting algorithm used in Section 5.5 is applied to the new adult data set. Except for the sampling rate of the sensors and the resulting window size, parameters (including features) will be kept identical to the pilot experiment. In particular, 441 features (63 features $\times$ 7 sensors = 441) are computed over 25,487 windows (representing approximately 3.5 hours of play data). Table 14 reports the average number of true positives, true negatives, false positives, and false negatives over fourteen play sessions as well as the number of positive and negative examples. Table 14 also provides the average recognition errors in terms of number of merged events, number of fragmented events, number of shortened events, and number of elongated events (see Section 3.2–Figure 3 and Appendix B for more details on these types of event–based errors). Table 15 reports the average performance of the toy–dependent models over fourteen play sessions according to 25 toy–independent categories. Figure 19 illustrates the distribution of F_1 scores grouped according to 25 toy–independent categories. Overall, the ensemble achieved an event–based F_1 score of $51.1\% \pm 16.4\%$ per toy–independent play activity. Table 16 compares performance of the models according event, segment, and time based evaluation criteria. Segment–based evaluation yields an F_1 score of $26.2\% \pm 8.7\%$ per toy–independent play activity, while time–based evaluation produces F_1 score of $39.2\% \pm 19.4\%$ per toy–independent play activity.

When compared to the pilot study, there were far fewer fragmentation errors resulting in higher event level accuracies. In particular, these boosted classifiers were more effective at classifying play events that experience abrupt changes in motion such as banging, shaking,

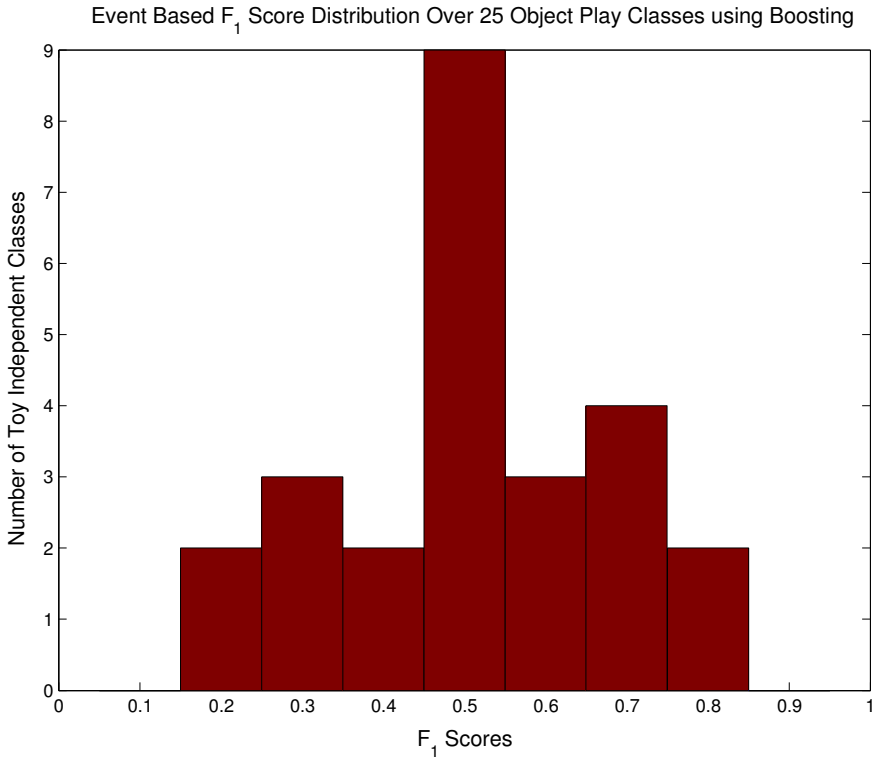


Figure 19: A histogram of event based F_1 scores for continuous classification using boosted 1–dimensional classifiers.

tossing the ball, thrashing, ramming, and hammering the puppy with the ring. The boosted classifiers were less effective with play events that exhibited more consistent periodic motions and aperiodic events, such as rocking, rolling, spinning, wearing the ring, and taking an imaginary drink from the dome. Definitions and examples of these play activities are listed in Appendix G.

6.5.2 Hidden Markov Models

Due to the temporal variations that can occur during a play session, Hidden Markov Models (HMMs) are one of the representations explored. Model inference was performed using the GT^2k (as described in Appendix C). For these experiments, various model topologies are explored, consisting of two to eight states. In particular, I will report on a three state, left–right topology where each state's observation probabilities were modeled using a mixture of two Gaussian distributions. Self transition probabilities were initialized to 60 percent with external transition probabilities initialized to 40 percent. A stochastic bi–gram grammar was

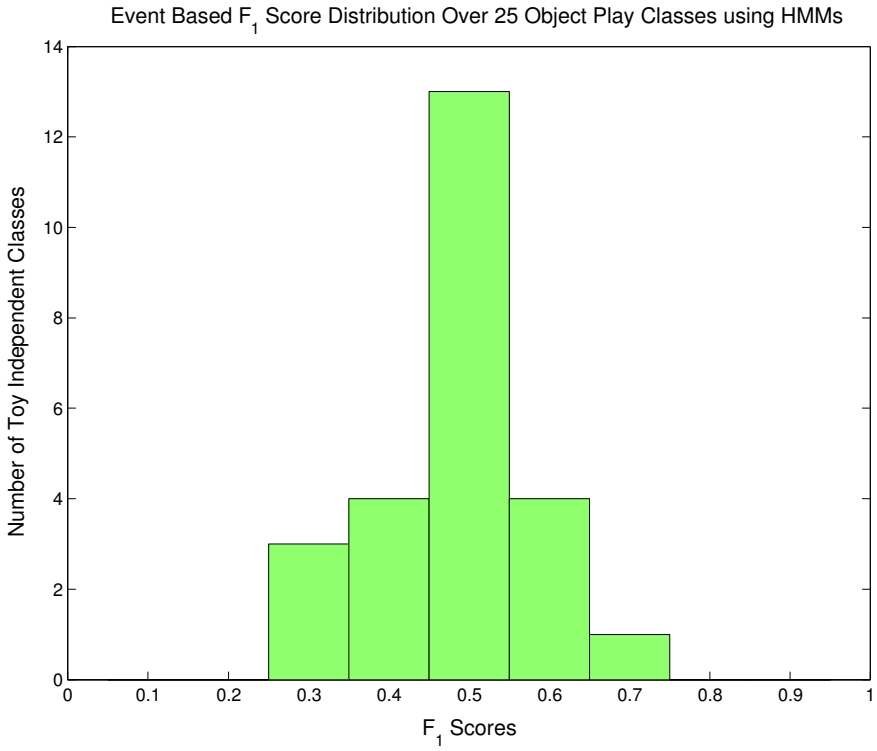


Figure 20: A histogram of event based F_1 scores for continuous classification using hidden Markov models.

constructed (based on 25 percent of the adult data set) and probabilistically applied during model alignment. Table 17 reports the overall performance of the toy–dependent models over fourteen play sessions trained using leave–one–out 13–fold cross validation. Figure 20 illustrates the distribution of F_1 scores grouped according to 25 toy–independent categories. Overall, the hidden Markov model achieved an event based F_1 score of $48.5\% \pm 9.3\%$ per toy–independent play activity.

As expected, the hidden Markov models were most effective at classifying play events that exhibited periodic motions. The models preformed equally well on play events containing both smooth and abrupt periodic motions. However, these models were less effective at identifying aperiodic motions, such as stacking blocks,[4] and periodic motions that contained decaying reverberations, such as rocking. Often, when an inverted dome is rocked, the toy will reverberate several times before an edge is tipped rocking the toy again.

[4]In this dissertation, stacking is considered to be aperiodic as each toy is only stacked once and the same motion is not happening repeatedly to the same toy.

6.5.3 Multiclass Support Vector Machines

Wang *et al.* use support vector machines to classify parent–infant social games within video sequences [70] (see Section 2.2.3). In a similar fashion, multiclass support vector machines can be applied to identify object play behavior. The *LIBSVM* software package is used to conduct several multiclass SVM recognition experiments [14].

All SVM experiments discussed in this section are conducted using a radial basis function (RBF) kernel. The cost, C , and gamma parameters, γ , are empirically selected by performing a search over fifty–six combinations of values using the 14 session subset of the adult data set. Five–fold cross validation is used to explore combinations $C = [.01, 100000]$ and $\gamma = [.001, 1000]$ in multiples of 10 increments. For feature vectors consisting of descriptive statistics and power spectral density features accuracy is maximized at 85.91% for $C = 100$ and $\gamma = 0.10$.

Once the model parameters were selected, experiments on the fourteen adult play session were constructed using leave–one–out cross validation using the same feature vectors as the previous two experiments. Table 18 reports the average performance of the toy–dependent models over the same fourteen play sessions used in the boosting and hidden Markov model experiments. Figure 21 illustrates the distribution of F_1 scores grouped according to 25 toy–independent categories.

Overall, the support vector machines achieved an event based F_1 score of $75.8\% \pm 24.7\%$ per toy–independent play activity. Table 19 compares performance of the models according event, segment, and time based evaluation criteria. Segment–based evaluation yields an F_1 score of $51.4\% \pm 23.1\%$ per toy–independent play activity, while time–based evaluation produces F_1 score of $50.6\% \pm 23.0\%$ per toy–independent play activity.

The support vector machines were effective at classifying a majority of both the periodic and aperiodic events. However, the classifier exhibited difficulties with aperiodic play events that had short temporal durations, such as spiking the puppy and stacking a Lego$^{\text{TM}}$ Quatro block on top of another toy. The classifier exhibited perfect retrieval on events that had very low within–class variation, such as wearing the ring, tossing the ball, sliding the toys on the ground, and having the puppy run to participant (listed as "runs towards self"). It

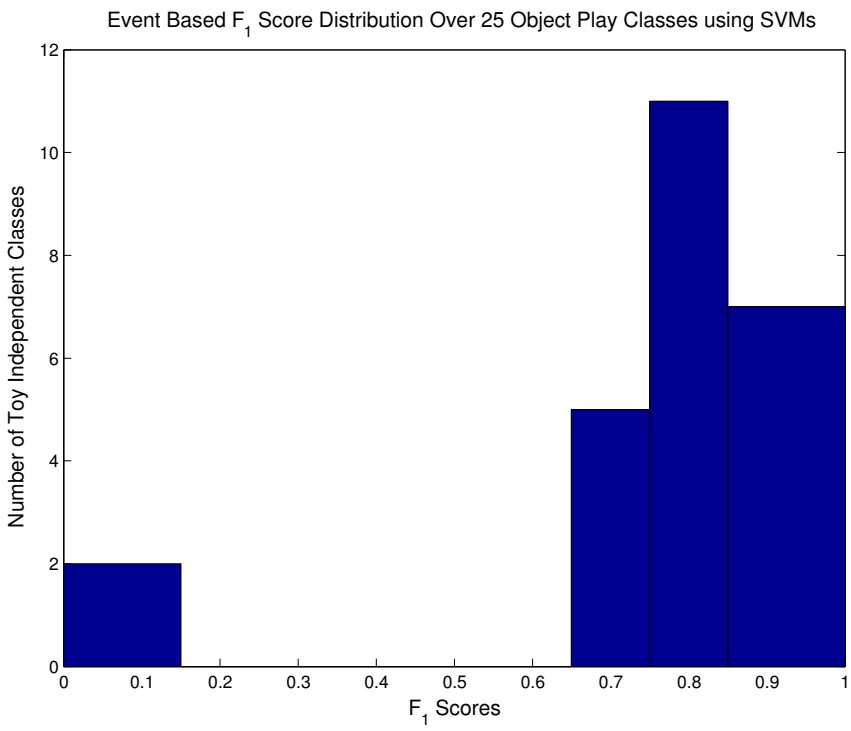


Figure 21: A histogram of event based F_1 scores for continuous classification using support vector machines.

is interesting to note, that while the participants were almost uniform in how "runs towards self" was performed, there was a much wider variation on how the plush puppy rattle was made to run around the play space for the "run" activity. This wider variation may account for the difference in effective retrieval scores between the two similar activities.

In comparisons to the boosted classifiers, the hidden Markov models, and the pilot experiments, the multiclass support vector machines showed increased performance in the overall F_1 score for event, segment, and time based evaluations. Figure 22 illustrates the distribution of the toy dependent event based F_1 scores for the support vector machines, hidden Markov models, and the boosted 1–dimensional classifier.

6.5.4 Generalization of Adult SVM models to Children's Play Data

Based on the results from the previous experiments, three support vector machines were trained on the full adult play data set and applied to the female child data set described in Section 6.2.2.2 for validation. Each SVM used a subset of the features described above

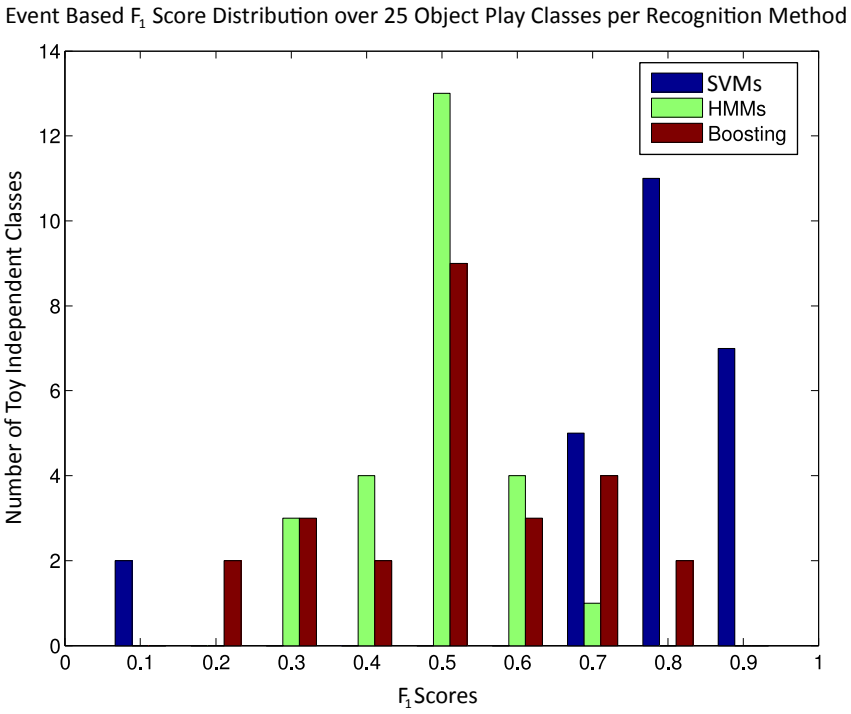


Figure 22: Histograms representing the toy–dependent event based F_1 scores for support vector machines, hidden Markov models, and boosted 1–dimensional classifiers.

and in Section 6.4.1 and were combined to help maximize generalization to the child's data set. The first SVM was trained using descriptive statistic–based features. The second SVM was trained using a combination of descriptive features and power spectral density features. The third was trained using correlative features and power spectral density features. The output of each model was combined forming a majority vote classifier. Table 20 reports the average number of true positives, true negatives, false positives, and false negatives over the child's four sessions as well as the number of positive and negative examples.

Table 21 lists event based performance metrics for the adult–trained SVM applied to the single child, four session data set. Overall, when applied to the child data the adult models achieved an average F_1 score of $58.8\% \pm 18.8\%$ per toy–independent play activity. Table 22 compares performance of the models according event, segment, and time based evaluation criteria. Segment–based evaluation yields an F_1 score of $14.8\% \pm 21.6\%$ per toy–independent play activity, while time–based evaluation produces F_1 score of $9.5\% \pm 15.9\%$ per toy–independent play activity.

6.6 Discussion

In this chapter I described the collection of multiple data sets and explored three different statistical techniques to model object play data in adults. While no one method dominated, support vector machines produced the higher F_1 score when compared to hidden Markov models and boosted decision stumps. By training models using combinations of three different feature spaces, the adult models were able to recognize a variety of play events from a single child over a four session data set. Although both the child and the adults were performing the same play protocol, the F_1 event score decreased when comparing performance on the adult and child data sets. This decrease is not unexpected as, upon visual inspection, the child data set had much wider within–class variation than the adult data set (see Chapter 8 for more details). The variations within a single play activity may also account for the much larger decrease in the segment and time based F_1 scores.

Despite these decreases, the performance of adult models on the child data is encouraging given the difficulty of collecting structured play data from young children. As will be discussed in Chapter 7, the adult models generalize well enough to support several aspects of retrospective review of play activities. Chapter 9 discusses the future application of these adult models towards other data sets discussed in this chapter.

Table 13: Occurrence of play primitives across all toys in the 4 play sessions of the female child multi-visit data set

Actions	Observed	Percent of Data	Duration in milliseconds		
			Minimum	Average	Maximum
NONE	572	23.3 %	35	339	2463
assemble	103	4.19 %	35	186	576
bang	41	1.67 %	35	161	470
breaks apart	19	0.77 %	24	55	109
carry	51	2.07 %	25	162	412
crush	42	1.71 %	20	178	579
drinks	9	0.37 %	101	132	164
drop	79	3.21 %	9	57	299
explore	54	2.20 %	40	308	1078
falls	23	0.93 %	22	73	135
flies	50	2.03 %	50	286	1047
hammers	6	0.24 %	98	145	193
ignore	7	0.28 %	25	465	1618
jumps	27	1.10 %	18	59	158
knocks down	5	0.20 %	33	62	84
nest	56	2.28 %	26	88	241
present	26	1.06 %	30	124	394
pushes	27	1.10 %	21	213	919
rams	23	0.93 %	21	134	401
relate	30	1.22 %	21	120	347
relocate	403	16.4 %	13	73	283
rock	26	1.06 %	39	215	514
roll	123	5.00 %	13	107	470
runs	44	1.79 %	7	107	531
runs towards self	20	0.81 %	40	170	817
separate	65	2.64 %	18	117	468
shake	50	2.03 %	23	111	411
slide	172	6.99 %	12	53	268
spikes	15	0.61 %	26	76	173
spin	22	0.89 %	39	169	340
stacks	14	0.57 %	25	172	868
thrashes	5	0.20 %	83	141	214
toss	22	0.89 %	25	119	582
touch	203	8.25 %	7	114	928
wears	26	1.06 %	81	508	3654

Total Play Events	1888
Total Overall Events	2460
Bias towards Play	77.0%
Bias towards Non–Play	23.3%

Table 14: Continuous recognition frequency statistics of events for boosted decision stumps over 14 play sessions

	Positives	Negatives	True Positives	False Positives	True Negatives	False Negatives	Merge	Fragmentation	Underfill	Overfill
non play	142	252	92	17	229	44	11	8	78	44
assemble	25	368	15	19	346	3	1	8	11	9
bang	9	387	6	10	362	3	0	1	7	1
drinks	3	395	1	7	368	2	0	1	1	1
explore	12	374	7	21	359	3	1	3	13	2
flies	8	392	5	5	363	2	0	2	8	2
hammers	2	402	2	1	369	1	0	1	2	1
jumps	8	392	3	5	363	5	0	1	3	2
nest	10	389	5	6	361	6	0	1	5	2
pushes	5	395	3	5	366	1	1	1	4	1
rams	6	395	4	3	365	3	0	1	4	1
relocate	49	313	31	45	322	16	1	3	15	14
rock	5	391	2	9	366	2	0	2	3	1
roll	16	374	11	17	355	3	1	3	12	3
runs	12	385	8	9	359	3	1	2	6	4
runstoward	2	399	1	3	369	2	0	0	1	0
separate	18	372	11	15	353	7	0	1	11	7
shake	10	390	8	4	361	3	1	1	10	2
slide	53	349	28	4	318	24	1	2	8	2
spikes	2	402	1	1	369	1	0	0	2	1
spin	5	390	3	11	366	2	1	1	4	1
stacks	2	402	1	1	369	1	0	1	1	1
thrashes	6	395	5	3	365	1	0	1	7	2
toss	2	401	2	1	369	1	0	0	2	1
wears	2	401	1	2	369	1	0	0	1	1

Table 15: Several metrics characterizing the event based performance of boosted decision stumps over 14 adult play sessions

	Accuracy	True Positive Rate (Recall)	False Positive Rate	Specificity	Positive Predicition Value (Precision)	Negative Prediction Value	Fale Discovery Rate	**F_1 Score**
non play	66.0%	64.8%	6.7%	93.1%	84.4%	83.9%	15.6%	73.3%
assemble	86.3%	60.0%	5.2%	94.8%	44.1%	99.1%	55.9%	50.8%
bang	89.6%	66.7%	2.6%	97.3%	37.5%	99.2%	62.5%	48.0%
drinks	90.5%	33.3%	1.8%	98.1%	12.5%	99.5%	87.5%	18.2%
explore	88.6%	58.3%	5.6%	94.5%	25.0%	99.2%	75.0%	35.0%
flies	90.3%	62.5%	1.3%	98.6%	50.0%	99.5%	50.0%	55.6%
hammers	91.3%	100.0%	0.2%	99.7%	66.7%	99.7%	33.3%	80.0%
jumps	89.0%	37.5%	1.3%	98.6%	37.5%	98.6%	62.5%	37.5%
nest	88.7%	50.0%	1.5%	98.4%	45.5%	98.4%	54.5%	47.6%
pushes	90.8%	60.0%	1.3%	98.7%	37.5%	99.7%	62.5%	46.2%
rams	90.5%	66.7%	0.8%	99.2%	57.1%	99.2%	42.9%	61.5%
relocate	80.7%	63.3%	14.4%	87.7%	40.8%	95.3%	59.2%	49.6%
rock	90.2%	40.0%	2.3%	97.6%	18.2%	99.5%	81.8%	25.0%
roll	88.7%	68.8%	4.5%	95.4%	39.3%	99.2%	60.7%	50.0%
runs	89.4%	66.7%	2.3%	97.6%	47.1%	99.2%	52.9%	55.2%
runstowarc	91.0%	50.0%	0.8%	99.2%	25.0%	99.5%	75.0%	33.3%
separate	87.7%	61.1%	4.0%	95.9%	42.3%	98.1%	57.7%	50.0%
shake	90.5%	80.0%	1.0%	98.9%	66.7%	99.2%	33.3%	72.7%
slide	79.1%	52.8%	1.1%	98.8%	87.5%	93.0%	12.5%	65.9%
spikes	91.1%	50.0%	0.2%	99.7%	50.0%	99.7%	50.0%	50.0%
spin	90.1%	60.0%	2.8%	97.1%	21.4%	99.5%	78.6%	31.6%
stacks	91.1%	50.0%	0.2%	99.7%	50.0%	99.7%	50.0%	50.0%
thrashes	91.3%	83.3%	0.8%	99.2%	62.5%	99.7%	37.5%	71.4%
toss	91.6%	100.0%	0.2%	99.7%	66.7%	99.7%	33.3%	80.0%
wears	91.1%	50.0%	0.5%	99.5%	33.3%	99.7%	66.7%	40.0%
Overall	88.2%	**61.4%**	2.5%	97.5%	**45.9%**	98.3%	54.1%	**51.1%**

Table 16: Comparison of overall performance of boosted decision stumps in terms of event, segment, and time based evaluations for a boosted classifier over 14 adult play sessions.

	Accuracy	True Positive Rate (Recall)	False Positive Rate	Specificity	Positive Predicition Value (Precision)	Negative Prediction Value	Fale Discovery Rate	**F_1 Score**
Events	88.2%	61.4%	2.5%	97.5%	45.9%	98.3%	54.1%	**51.1%**
Time (ms)	92.5%	38.3%	2.3%	97.7%	42.3%	97.7%	57.7%	**39.2%**
Segments	88.6%	26.8%	3.4%	96.6%	28.2%	96.6%	71.8%	**26.2%**

Table 17: Several metrics characterizing the event based performance of continuous recognition using hidden Markov models over 14 adult play session.

	Accuracy	True Positive Rate (Recall)	False Positive Rate	Specificity	Positive Predicition Value (Precision)	Negative Prediction Value	Fale Discovery Rate	**F_1 Score**
assemble	86.2%	52.4%	4.2%	95.8%	46.0%	96.7%	54.0%	49.0%
bang	93.1%	46.2%	1.6%	98.4%	50.9%	98.0%	49.1%	48.4%
drinks	98.3%	58.3%	0.7%	99.3%	23.3%	99.8%	76.7%	33.3%
explore	89.0%	38.7%	2.3%	97.7%	49.6%	96.5%	50.4%	43.5%
flies	94.1%	50.0%	1.3%	98.7%	57.7%	98.3%	42.3%	53.6%
hammers	98.9%	52.9%	0.3%	99.7%	45.0%	99.8%	55.0%	48.6%
jumps	95.0%	59.2%	1.6%	98.4%	44.2%	99.1%	55.8%	50.6%
nest	87.1%	29.0%	1.8%	98.2%	52.9%	95.1%	47.1%	37.4%
pushes	96.2%	41.3%	0.8%	99.2%	50.0%	98.9%	50.0%	45.2%
rams	94.0%	33.7%	1.1%	98.9%	48.5%	98.0%	51.5%	39.8%
relocate	64.8%	55.0%	11.2%	88.8%	53.6%	89.4%	46.4%	54.3%
rock	95.7%	36.0%	1.2%	98.8%	31.0%	99.0%	69.0%	33.3%
roll	89.8%	58.4%	2.7%	97.3%	58.1%	97.3%	41.9%	58.3%
runs	93.4%	61.0%	2.2%	97.8%	46.2%	98.8%	53.8%	52.6%
runstoward	98.4%	46.3%	0.2%	99.8%	79.2%	99.3%	20.8%	58.5%
separate	86.4%	51.4%	4.5%	95.5%	39.0%	97.2%	61.0%	44.3%
shake	89.7%	43.4%	1.6%	98.4%	64.8%	96.2%	35.2%	52.0%
slide	57.3%	50.9%	14.7%	85.3%	45.2%	87.9%	54.8%	47.9%
spikes	99.3%	60.0%	0.1%	99.9%	75.0%	99.8%	25.0%	66.7%
spin	97.5%	64.6%	0.7%	99.3%	56.4%	99.5%	43.6%	60.2%
stacks	98.8%	28.6%	0.3%	99.7%	28.6%	99.7%	71.4%	28.6%
thrashes	96.1%	56.0%	1.3%	98.7%	39.4%	99.3%	60.6%	46.3%
toss	99.0%	55.0%	0.2%	99.8%	61.1%	99.7%	38.9%	57.9%
wears	99.3%	63.6%	0.2%	99.8%	46.7%	99.9%	53.3%	53.8%
Overall	91.5%	**49.7%**	2.4%	97.6%	**49.7%**	97.6%	50.3%	**48.5%**

Table 18: Several metrics characterizing the event based performance of support vector machines over 14 adult play session.

	Accuracy	True Positive Rate (Recall)	False Positive Rate	Specificity	Positive Predicition Value (Precision)	Negative Prediction Value	Fale Discovery Rate	**F_1 Score**
non play	86.4%	84.4%	2.1%	97.7%	94.2%	97.3%	5.8%	89.0%
assemble	89.6%	84.2%	3.4%	96.6%	84.2%	96.6%	15.8%	84.2%
bang	92.3%	79.2%	1.3%	98.6%	82.6%	98.3%	17.4%	80.9%
drinks	93.6%	66.7%	0.6%	99.3%	75.0%	99.0%	25.0%	70.6%
explore	90.3%	82.2%	2.9%	97.1%	82.2%	97.1%	17.8%	82.2%
flies	93.2%	81.8%	1.3%	98.6%	81.8%	98.6%	18.2%	81.8%
hammers	95.1%	75.0%	0.3%	99.7%	75.0%	99.7%	25.0%	75.0%
jumps	95.1%	80.0%	0.3%	99.7%	80.0%	99.7%	20.0%	80.0%
nest	94.2%	71.4%	0.6%	99.3%	71.4%	99.3%	28.6%	71.4%
pushes	94.5%	80.0%	0.6%	99.3%	80.0%	99.3%	20.0%	80.0%
rams	95.1%	80.0%	0.3%	99.7%	80.0%	99.7%	20.0%	80.0%
relocate	93.8%	76.9%	1.0%	99.0%	76.9%	99.0%	23.1%	76.9%
rock	93.8%	78.6%	0.6%	99.3%	84.6%	99.0%	15.4%	81.5%
roll	94.4%	77.8%	0.6%	99.3%	77.8%	99.3%	22.2%	77.8%
runs	95.1%	66.7%	0.3%	99.7%	66.7%	99.7%	33.3%	66.7%
runstowarc	95.7%	100.0%	0.0%	100.0%	100.0%	99.7%	0.0%	100.0%
separate	95.4%	100.0%	0.3%	99.7%	66.7%	99.7%	33.3%	80.0%
shake	95.1%	100.0%	0.3%	99.7%	50.0%	99.7%	50.0%	66.7%
slide	95.4%	100.0%	0.0%	100.0%	100.0%	99.7%	0.0%	100.0%
spikes	94.8%	0.0%	0.3%	99.7%	0.0%	99.7%	100.0%	0.0%
spin	95.1%	100.0%	0.3%	99.7%	75.0%	99.7%	25.0%	85.7%
stacks	94.8%	0.0%	0.3%	99.7%	0.0%	99.7%	100.0%	0.0%
thrashes	95.1%	100.0%	0.3%	99.7%	75.0%	99.7%	25.0%	85.7%
toss	95.7%	100.0%	0.0%	100.0%	100.0%	100.0%	0.0%	100.0%
wears	95.4%	100.0%	0.0%	100.0%	100.0%	99.7%	0.0%	100.0%
Overall	94.0%	**78.6%**	0.7%	99.2%	**74.4%**	99.2%	25.6%	**75.8%**

Table 19: Comparison of overall performance in terms of event, segment, and time based evaluations for SVMs.

	Accuracy	True Positive Rate (Recall)	False Positive Rate	Specificity	Positive Predicition Value (Precision)	Negative Prediction Value	Fale Discovery Rate	**F_1 Score**
Events	94.0%	78.6%	0.7%	99.2%	74.4%	99.2%	25.6%	75.8%
Time (ms)	97.1%	48.8%	1.6%	98.4%	53.2%	94.5%	42.8%	50.6%
Segments	93.2%	60.0%	2.4%	97.6%	51.2%	97.6%	44.8%	51.4%

Table 20: Recognition frequency statistics of the adult–trained majority vote SVMs applied to the child data

	Positives	Negatives	True Positives	False Positives	True Negatives	False Negatives	Merge	Fragmentation	Underfill	Overfill
non play	53	135	49	7	135	3	11	2	9	9
assemble	10	180	5	4	179	5	1	0	1	1
bang	4	188	3	2	185	2	0	0	3	1
breaksapar	2	191	1	0	187	2	0	0	0	0
carry	4	189	2	0	184	3	0	0	1	0
crush	3	190	2	1	185	2	0	0	0	0
drinks	1	192	1	1	187	1	1	0	1	0
drop	6	187	3	0	182	4	0	0	1	0
explore	4	186	1	4	184	3	1	1	2	1
falls	2	191	1	0	186	1	0	0	1	0
flies	5	189	2	1	184	3	0	1	2	0
hammers	1	192	1	1	188	1	1	0	1	1
jumps	2	188	1	3	186	2	0	0	1	1
knockdowr	1	192	1	0	188	1	0	0	0	0
nest	5	189	1	0	184	4	0	0	0	0
present	2	191	1	0	186	2	0	0	0	0
pushes	2	191	1	0	186	2	0	0	0	0
rams	2	191	1	0	187	2	0	0	0	0
relocate	28	165	11	0	160	18	0	1	1	0
rock	2	186	1	5	186	1	1	1	1	1
roll	10	182	4	2	179	5	1	1	3	1
runs	3	190	1	1	185	3	0	1	1	0
runstoward	2	191	1	0	187	2	0	0	0	0
separate	6	187	2	0	183	4	0	0	1	0
shake	4	189	3	1	184	2	0	1	3	1
slide	13	181	4	0	176	9	0	1	1	0
spikes	2	191	1	1	187	2	0	0	0	0
spin	2	190	1	2	187	1	0	1	1	1
stacks	2	191	1	0	187	1	0	0	0	0
thrashes	1	192	1	0	188	1	0	0	0	0
toss	2	190	1	2	187	1	0	1	1	1
touch	17	177	7	0	172	10	0	1	1	0
wears	3	191	1	0	186	2	0	1	0	0

Table 21: Metrics characterizing the event based performance of the adult trained SVMs applied to the child data

	Accuracy	True Positive Rate (Recall)	False Positive Rate	Specificity	Positive Predicition Value (Precision)	Negative Prediction Value	Fale Discovery Rate	**F_1 Score**
non play	92.6%	92.5%	5.2%	95.1%	87.5%	97.8%	12.5%	89.9%
assemble	92.1%	50.0%	2.2%	97.8%	55.6%	97.3%	44.4%	52.6%
bang	95.8%	75.0%	1.1%	98.9%	60.0%	98.9%	40.0%	66.7%
breaksapar	96.4%	50.0%	0.0%	100.0%	100.0%	98.9%	0.0%	66.7%
carry	94.8%	50.0%	0.0%	100.0%	100.0%	98.4%	0.0%	66.7%
crush	95.3%	66.7%	0.5%	99.5%	66.7%	98.9%	33.3%	66.7%
drinks	96.4%	100.0%	0.5%	99.5%	50.0%	99.5%	50.0%	66.7%
drop	93.8%	50.0%	0.0%	100.0%	100.0%	97.8%	0.0%	66.7%
explore	93.7%	25.0%	2.2%	97.9%	20.0%	98.4%	80.0%	22.2%
falls	96.4%	50.0%	0.0%	100.0%	100.0%	99.5%	0.0%	66.7%
flies	93.8%	40.0%	0.5%	99.5%	66.7%	98.4%	33.3%	50.0%
hammers	96.9%	100.0%	0.5%	99.5%	50.0%	99.5%	50.0%	66.7%
jumps	95.8%	50.0%	1.6%	98.4%	25.0%	98.9%	75.0%	33.3%
knockdowr	97.4%	100.0%	0.0%	100.0%	100.0%	99.5%	0.0%	100.0%
nest	93.3%	20.0%	0.0%	100.0%	100.0%	97.9%	0.0%	33.3%
present	95.9%	50.0%	0.0%	100.0%	100.0%	98.9%	0.0%	66.7%
pushes	95.9%	50.0%	0.0%	100.0%	100.0%	98.9%	0.0%	66.7%
rams	96.4%	50.0%	0.0%	100.0%	100.0%	98.9%	0.0%	66.7%
relocate	79.3%	39.3%	0.0%	100.0%	100.0%	89.9%	0.0%	56.4%
rock	96.3%	50.0%	2.7%	97.4%	16.7%	99.5%	83.3%	25.0%
roll	91.7%	40.0%	1.1%	98.9%	66.7%	97.3%	33.3%	50.0%
runs	94.3%	33.3%	0.5%	99.5%	50.0%	98.4%	50.0%	40.0%
runstowarc	96.4%	50.0%	0.0%	100.0%	100.0%	98.9%	0.0%	66.7%
separate	93.8%	33.3%	0.0%	100.0%	100.0%	97.9%	0.0%	50.0%
shake	95.3%	75.0%	0.5%	99.5%	75.0%	98.9%	25.0%	75.0%
slide	88.1%	30.8%	0.0%	100.0%	100.0%	95.1%	0.0%	47.1%
spikes	95.9%	50.0%	0.5%	99.5%	50.0%	98.9%	50.0%	50.0%
spin	96.4%	50.0%	1.1%	98.9%	33.3%	99.5%	66.7%	40.0%
stacks	96.9%	50.0%	0.0%	100.0%	100.0%	99.5%	0.0%	66.7%
thrashes	97.4%	100.0%	0.0%	100.0%	100.0%	99.5%	0.0%	100.0%
toss	96.4%	50.0%	1.1%	98.9%	33.3%	99.5%	66.7%	40.0%
touch	87.1%	41.2%	0.0%	100.0%	100.0%	94.5%	0.0%	58.3%
wears	95.4%	33.3%	0.0%	100.0%	100.0%	98.9%	0.0%	50.0%
Overall	94.4%	**55.7%**	0.8%	99.2%	**74.6%**	98.3%	25.4%	**58.8%**

Table 22: Comparison of overall performance in terms of event, segment, and time based evaluations for SVMs trained on adult data and applied to the child data.

	Accuracy	True Positive Rate (Recall)	False Positive Rate	Specificity	Positive Predicition Value (Precision)	Negative Prediction Value	Fale Discovery Rate	**F_1 Score**
Events	94.4%	55.7%	0.8%	99.2%	74.6%	98.3%	25.4%	58.8%
Time (ms)	92.9%	9.3%	3.9%	96.1%	16.6%	97.5%	35.6%	9.5%
Segments	89.6%	19.7%	4.0%	96.0%	14.2%	96.7%	38.0%	14.8%

CHAPTER VII

ACCEPTABLE RECOGNITION RATES FOR RETROSPECTIVE ANALYSIS OF CHILDREN'S PLAY BEHAVIORS

The study outlined in this chapter is designed to determine the quality of automatic recognition required to support retrospective review of children's play activities. The results of this study will help inform which recognition algorithms (and associated accuracies) are acceptable for use in future research and applications involving the *Child'sPlay* system.

This chapter begins with the research questions addressed by this study and my hypothesis (Section 7.1). Afterwards, I present the within–subjects study design (Section 7.3) which includes a list of the conditions, the method, and the selection criteria for participants. The chapter then discusses the data collected (Section 7.4), the subsequent analysis (Section 7.5), and the overall findings from this study.

7.1 Research Questions and Hypothesis

The goal of this study is to gain a better understanding of the number and types of recognition errors that a user can tolerate while annotating or reviewing previously captured object play data. This study is designed to address the following Research Question 3 and sub questions:

1. What level of recognition quality does a user find acceptable when using an interface for retrospective review?

2. What is the perceived effort to correctly identify object play activities relative to the level of provided annotation quality?

3. How is task performance impacted by recognition quality?

In particular, I hypothesize that there is a certain level of recognition quality that users are willing to tolerate when using a visualization to search for information. Even in the

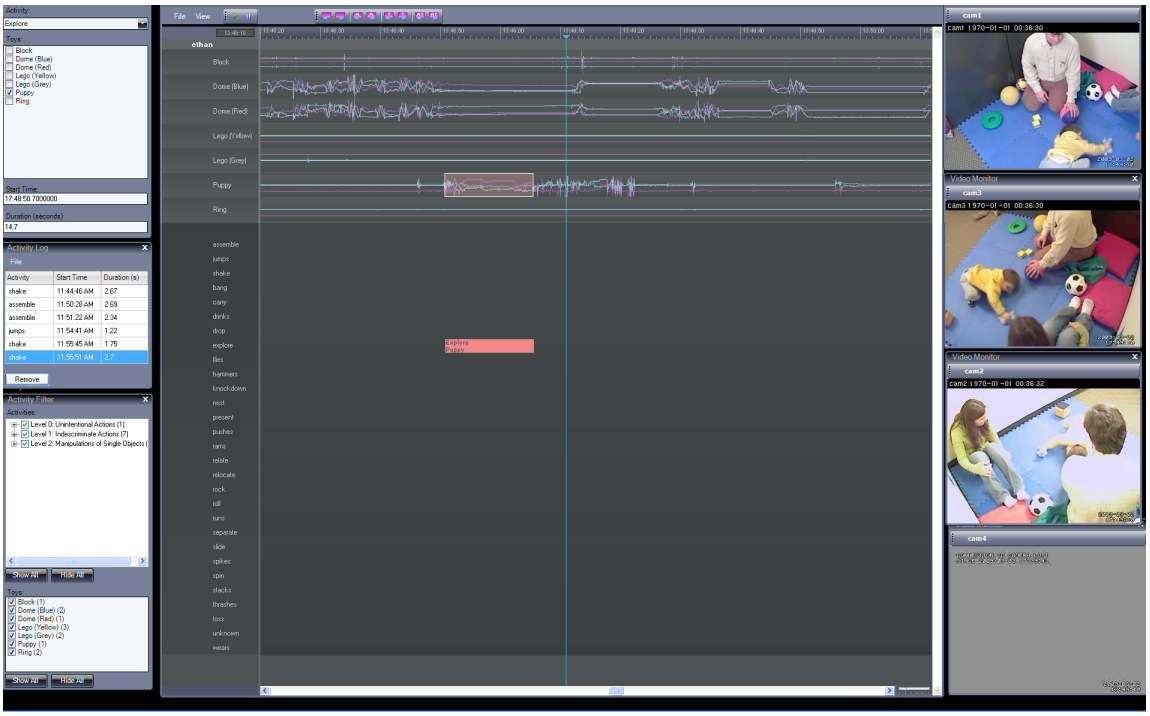


Figure 23: Screen capture of the *PlayView* interface

presence of errors, automatic recognition can help increase the percentage of video reviewed, increase the number of play activities identified, and reduce the perceived effort to identify object play behaviors.

7.2 Interface for Retrospective Review

The *PlayView* interface used by the data coders to annotate the adult and children data sets (as discussed in Section 6.2.1) is used to evaluate the impact that the quality of computer supplied annotations has on the task of retrospective review. The *PlayView* interface is used to view object play data and create annotations of the play data. The *PlayView* interface supports the display of video from multiple camera views, the display of audio signatures, the display of accelerometer signatures, the selective display of user created and computer generated play annotations, as well as the ability to search for specific play events. This section will briefly describe the main components of the *PlayView* interface as it relates to this study.

7.2.1 Videos Windows and the Timeline

The *PlayView* interface supports viewing of multiple video feeds. For the purposes of this study, video from four cameras will be displayed along the right hand side of the interface in four video windows. The windows can be enlarged to expose more detail and rearranged to suit a user's preference. Across the top of the interface, in the center panel, is the time line. Time moves forward from left to right. The play head is the blue horizontal line, and it indicates where the video views are in the relation to the timeline. In Figure 23, the play head is positioned at 13:49:10 on the timeline.

7.2.2 Toy View

Below the time line is the annotation view area. The top most section is the *Toy View*. In this view motion activity is listed per toy. The toys are listed alphabetically. When activity occurs on a toy, a visualization of the acceleration data is graphed in the toy's track in the appropriate area under the timeline. The accelerometer traces in Figure 23 show that the dome toys (formed as the ball) experience motion as the father rolls the ball while the child plays with the puppy. Both of the Lego^{TM} Quatro toys are not experiencing motion during this time. [1]

Figure 23 has a single pink annotation indicating that the plush puppy rattle toy is being explored. Annotations can be drawn in the toy view by clicking and dragging on the track of the toy which is experiencing motion. When an annotation is drawn in the toy view, the type of activity must be selected from the annotation property pane (in the upper left hand corner) or from a context menu. Once an annotations is assigned an activity, the annotations also appears in the activity view.

7.2.3 Activity View

The *Activity View* is located below the toy view. Annotations in this view are categorized by activity and temporally (vertically) correspond to labels which appear in the toy view.

[1] For the purposes of the study, participants do not have access to the accelerometer visualizations to prevent them from using the traces to identify play behaviors (instead of, or, in addition to the computer supplied annotations). The activity filter is also excluded for similar reasons.

Annotations are created in the activity view in a similar fashion to the toy view. Except, when an annotation is drawn in an activity track, the toys involved must be selected from the activity pane. Once an annotation on an activity track is assigned toys, the annotations also appear in the toy view.

7.2.4 Properties Pane

The annotation *Property Pane* shows information on the annotation currently selected. Figure 23 the pink annotation highlighting exploration of the puppy is currently selected. In the properties pane the activity "Explore" appears in a list box, and "puppy" is checked. The properties pane includes the starting time of the annotation as well as the duration of the annotation.

7.2.5 Activity Log

The activity log appears beneath the property pane. This element was added specifically for study participants and not used by the data coders. The Log window keeps track of annotations that study participants have viewed and wish to record as play events.

7.3 Study Design

This study is a within–subjects design with 20 participants. There are four conditions that are explored in this study to help assess the effect of computer recognition capabilities on the task of retrospective review. These conditions vary by the quality of automatic play–recognition support supplied to the user via the *PlayView* interface. In the control condition, participants are asked to identify three play behaviors using the interface with no computer supplied annotations. In the other three conditions participants are asked to identify the same three play behaviors using the interface with computer generated annotations containing a range of low, medium, and high effective retrieval rates. Each participant receives 30 minutes of training on both proper annotation of the three object play activities for which they were searching (10 minutes) and interface software usage (20 minutes). Each participant is asked to identify the number of occurrences for each of the three play task within the entire video, remove instances where the computer incorrectly

identified one of the three events, and ensure that annotations identifying play activities correctly represent the start and end of the play activity. After each condition, participants complete a NASA–TLX survey, as well as a post–condition questionnaire to ascertain the perceived workload and level of frustration in relation to the quality of the automatic assistance provided and the ease of completing a given task.

7.3.1 Conditions

In each condition, participants concurrently search through one of four recorded play sessions for occurrences of the three play behaviors listed below. These three behaviors were selected because they span multiple levels of play sophistication and were independent from statistical model development.

1. **shaking any toy:** any of the seven toys being shaken
   (Level 1: Indiscriminate Actions)

2. **assembling Lego™ Quatros:** assembling the two Lego™ Quatro toys
   (Level 4: Presentation and General Combinations)

3. **puppy jumping:** the plush puppy rattle jumping over any of the other six toys
   (Level 5: Object Directed)

During the course of the experiment, participants search through play data collected from four different play sessions involving a 5–year–old child performing the adult play protocol (described in Section 6.2). Participants search through a different play session for each condition. The presentation order of the play sessions is held constant across all 20 participants as all the play sessions involve the same child and are similar in both length as well as play behavior frequency. On average, there are 27 instances of the play behaviors that the participants should identify. To minimize the ordering effect, the order of conditions is balanced using a partial Latin Square. However, participants remain blind to the order in which they receive the four conditions: None, Motion–Only, Low, and High.

7.3.1.1 None

In the None condition participants receive no recognition assistance within the *PlayView* interface. The toy and activity views contain no additional information to support the search process. In this condition, participants can watch the video in real–time or use the interface quickly scan through video. Participants must create all annotations. This condition is the control condition of the experiment because it is similar to current best practices in retrospective analysis. Figure 24 is a screen capture of this condition applied to the fourth play session.

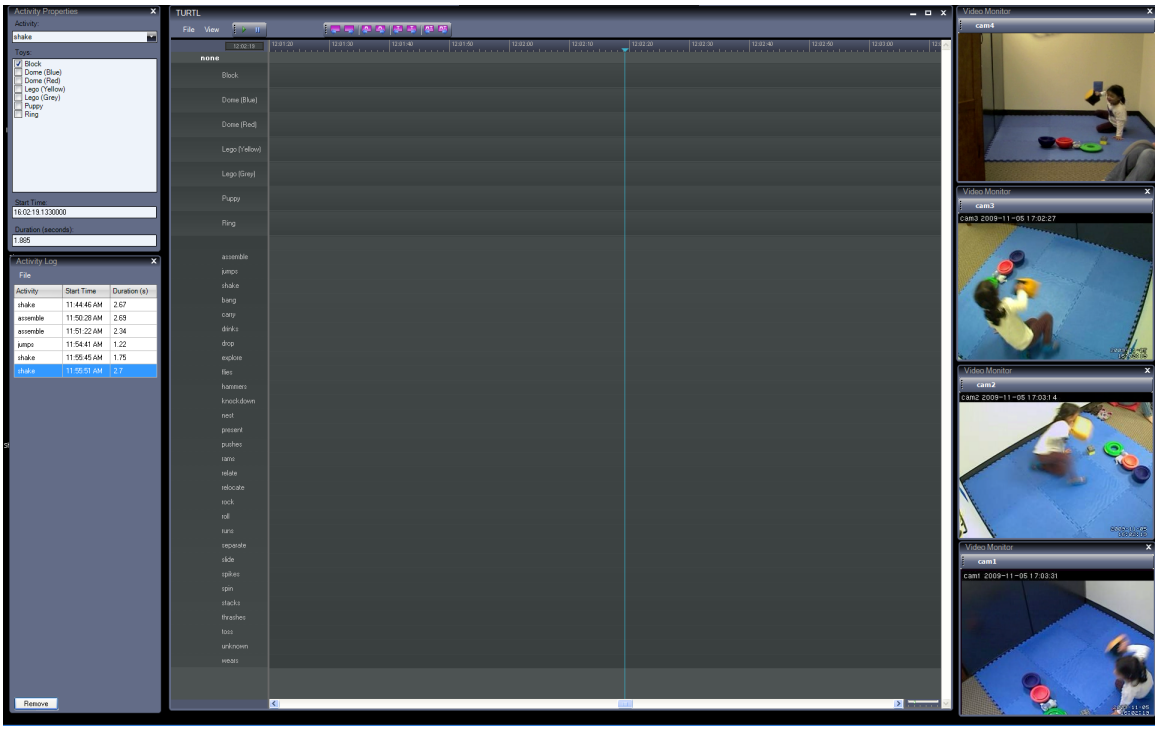


Figure 24: Screen capture of the *PlayView* interface while searching through the fourth play session during the None condition.

7.3.1.2 Motion–Only

Participants in the Motion–Only condition receive a naïve level of recognition assistance within the *PlayView* interface. In this condition, every instance of motion experienced by the toys is annotated as a "unknown" activity and highlighted grey in color. Participants can use the toy and activity views to visually correlate when toys are experiencing motion with

the video time line. Participants can quickly jump the video over spaces where no motion is occurring, quickly scan through video, or watch the video in real–time. Participants can annotate data by changing the categorization of the "unknown" annotations to more descriptive labels, or they can create a new annotation. Figure 25 is a screen capture of this condition applied to the fourth play session.

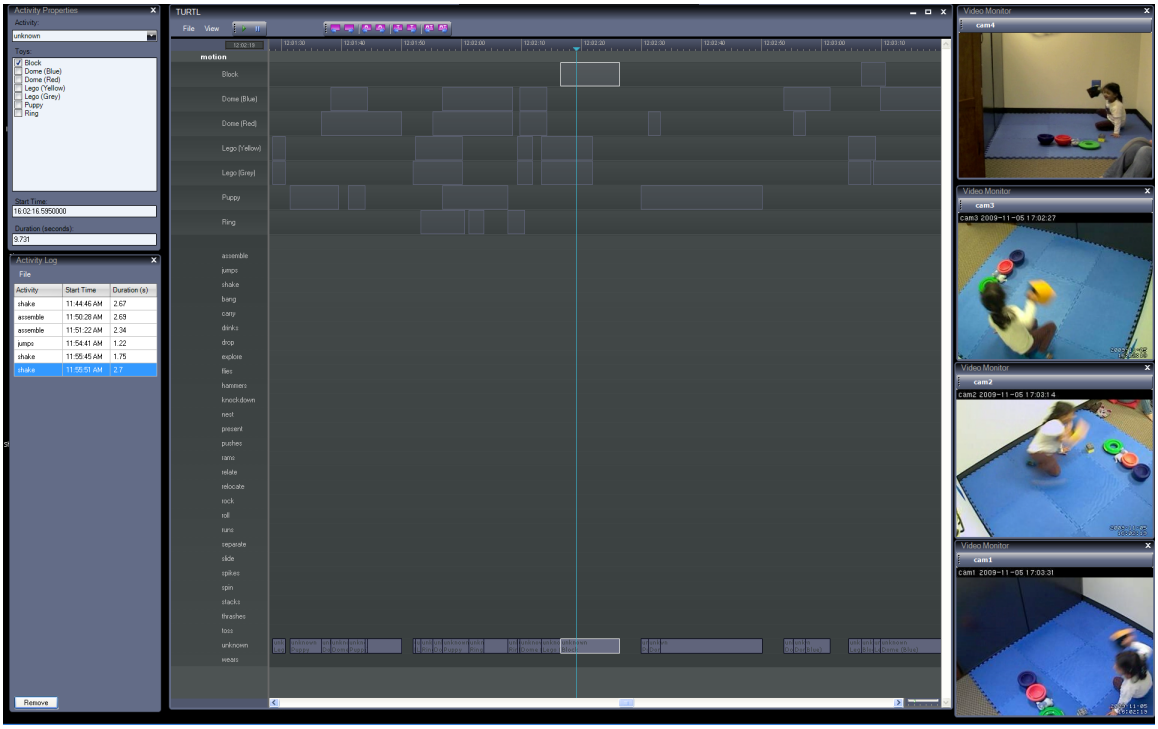


Figure 25: Screen capture of the *PlayView* interface while searching through the fourth play session during the MOTION–ONLY condition.

7.3.1.3 LOW

Participants in the LOW condition receive recognition assistance within the *PlayView* interface. In this condition, 78 toy–dependent play activities are annotated with bright colors using the statistical models described in Section 6.5.4, and the supplied annotations reflect the current capability of the *Child'sPlay* system. Participants can use the toy and activity view to visually correlate when toys are experiencing motion with the video time line. Participants can quickly jump to specific activities within the video, jump past segments of the video where no motion is occurring, quickly scan through video, or watch the video

in real–time. Participants can annotate data by adjusting the boundaries of existing labels, changing the categorization of the other annotations, or can create a new annotation. For consistency, the MOTION–ONLY grey labels are also present in this condition with the higher level recognition information superimposed. In this condition, recognition support provided for each of the three play activities experiences a different type of recognition error. Lego^{TM} Quatro activities experienced high insertion errors, puppy jumping experienced a high number of deletion errors, and shaking toys was often misidentified as banging. Using the current recognition capabilities discussed in Section 6.5.4 the average F_1 measure of the provided annotations is 58.8%. The average F_1 measure for the three activities the participants must identify in this condition is 47.38%. Figure 26 is a screen capture of this condition applied to the fourth play session.

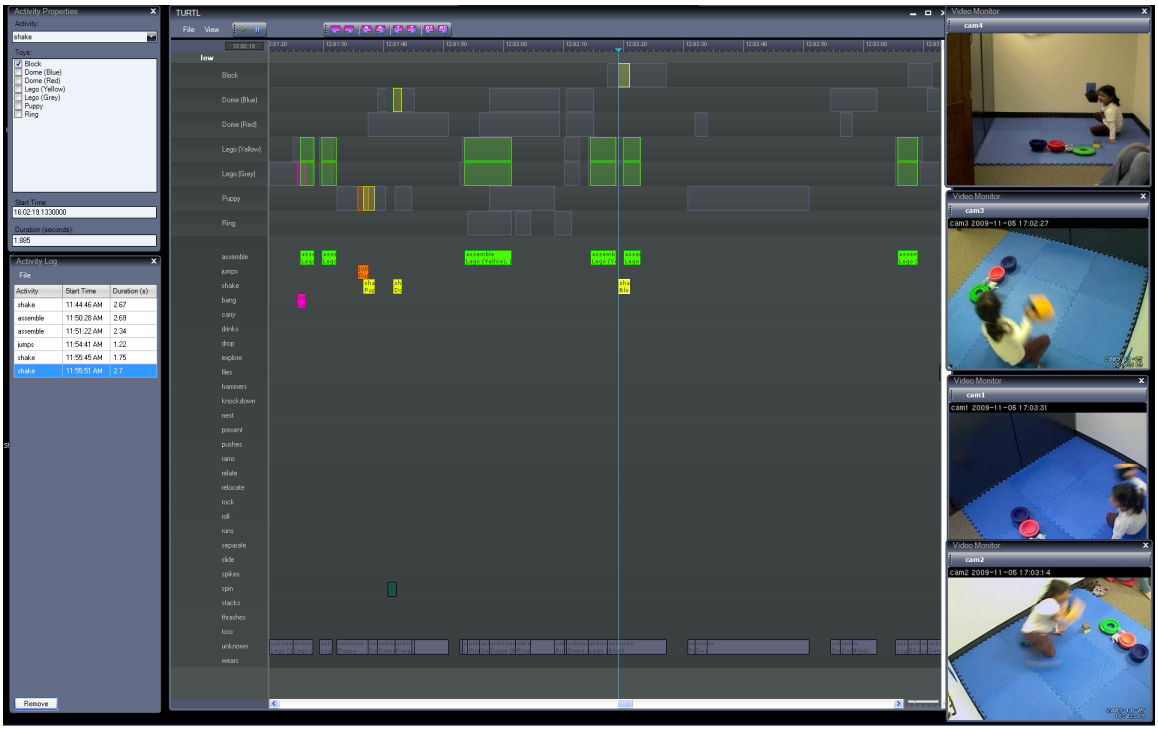


Figure 26: Screen capture of the *PlayView* interface while searching through the fourth play session during the LOW condition.

7.3.1.4 HIGH

Participants in the HIGH condition receive the highest quality of recognition assistance within the *PlayView* interface. Similar to the LOW condition, 78 toy–dependent play activities are annotated with bright colors using the statistical models described in Section 6.5.4. However, modifications were made by hand to improve the supplied annotations to significantly increase the F_1 measure. This condition is intended to represent the future capability of the *Child'sPlay* system. Identical to the LOW condition, participants can use the toy and activity view to visually correlate when toys are experiencing motion with the video time line. Participants can quickly jump to specific activities within the video, jump past segments of the video where no motion is occurring, quickly scan through video, or watch the video in real–time. Participants can annotate data by adjusting the boundaries of existing labels, changing the categorization of the other annotations, or can create a new annotation. For consistency, the MOTION–ONLY grey labels are also present in this condition with the higher level recognition information superimposed. In this condition, recognition support provided for each of the three play activities reduces the errors experienced in the LOW condition. Insertion errors for the Lego^{TM} Quatro activities are reduced by 65%, deletion errors for the puppy jumping are reduced 65%, and substitution errors between shaking and banging toys is also reduced by 65%. The average F_1 measure of the three activities to identify in this condition is 68.50%. Figure 27 is a screen capture of this condition applied to data collected in the fourth play session.

7.3.2 Method

7.3.2.1 Informed Consent and Background Survey

After a standard informed consent procedure, the participant completes a background survey to collect basic demographic information, as well as experience using video editing and annotation software; information on exposure to pattern recognition courses; and overall trust in information automatically provided by computer algorithms is also ascertained. Background data is also collected on overall experience with computers, involvement in clinical studies observing play, and exposure to children, in general. Appendix E includes

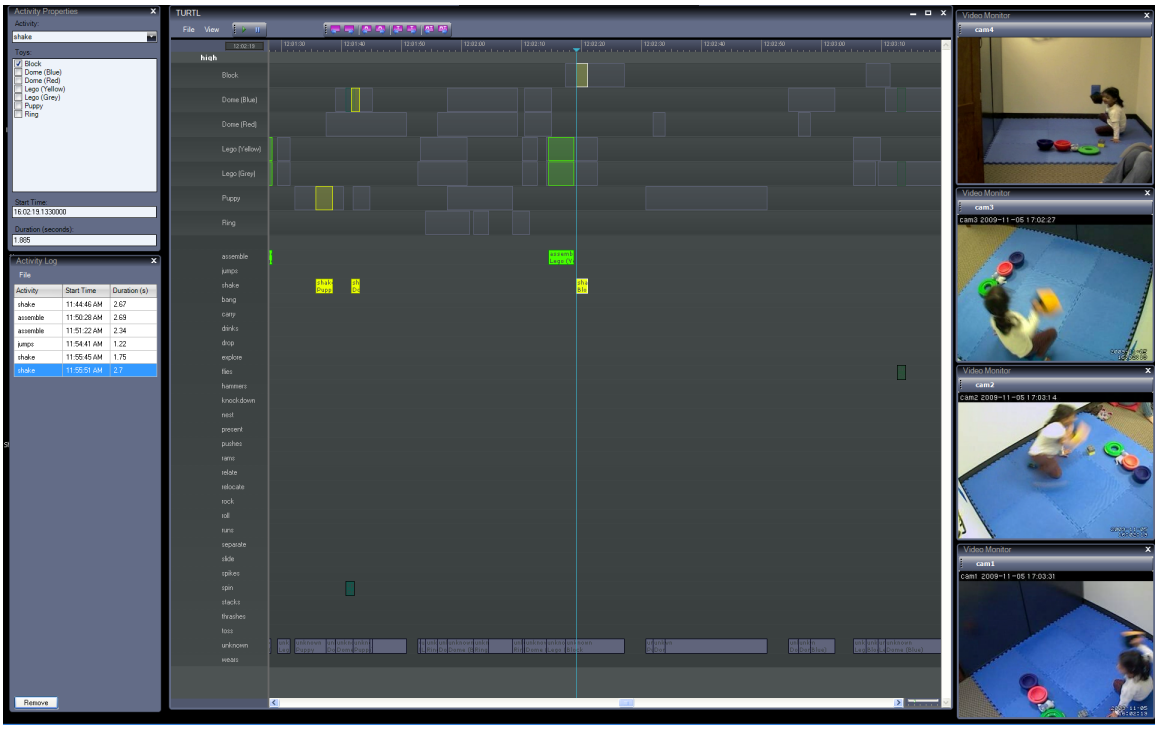


Figure 27: Screen capture of the *PlayView* interface while searching through the fourth play session during the HIGH condition.

the background survey as well as all other paper–based surveys administered during the experiment.

7.3.2.2 Training

After completing the background survey, the participant is informed of the task he will be completing. Namely, he will be viewing four different videos of a child playing and identifying when three specific types of play occur. However, the types of play are not described at this time. Next, the participant is physically shown the seven toys used during play. Each toy is identified by type and color. Afterward, the sensor inside the yellow Lego^{TM} Quatro is revealed. The participant is told that there are similar sensors in each of the toys and that the toys transmit motion information back to the computer. The computer then analyzes the motion and identifies the type of play that is occurring to the best of its ability.

Using the physical toys, the participant is shown each of the three play activities he will be identifying in the video. First the participant is shown the two Lego^{TM} Quatro toys being

assembled. He is instructed that the order of assembly does not matter, and that the child is not required to physically hold both toys. The main requirement is that the one–four pegs of one Lego^{TM} Quatro are mated into the underside holes of the other Lego^{TM} Quatro. Second, the participant is shown the plush puppy rattle toy jumping over another toy. He is instructed that it does not matter which toy the puppy jumps over, the number of toys traversed, nor the height of the jump. The only requirement is that the puppy starts on the ground, passes over a toy while in the air and returns back to the ground. Third, the participant is shown each of the toys being shaken, individually. He is instructed that the direction of the shake, the toy orientation, and the frequency do not matter. After seeing a description of the play activities the participant is told that all play activities begin the moment one of the involved toys is touched and ends when the last toy involved comes to rest (be it in a statically held position, or resting on the ground). Questions about the play activities are then fielded. Explanation of these activities and related questions typically lasts 10 minutes.

Once the participant expresses he is comfortable and feels he can identify the three play activities, he is told of his primary task and trained in usage of the *PlayView* interface. Each participant is told that he must identify as many occurrences of the three play tasks as possible and reach the end of the video. They are also to remove annotations where the computer incorrectly identifies any of the three events and ensure that annotations correctly represent the start and end of the play activity.

All participants received training on the same play session, which was used solely during the training session. The participants are trained using a 20 minute play session from the adult play data set that is annotated with both motion labels and high recognition. First, each section of the interface is named and described, starting with the videos, the toy view, the activity view, the properties pane, and then the log window. Next the timeline is described, and basic functionality of the play–head is explained. Participants are shown how to scan through the video by dragging the play–head, jump to various points in the time line by double clicking on it, and play the video in real–time. After instruction, the participant is asked to repeat the activities he has just been taught.

Next, the participant is shown how to identify computer supplied annotations. As part of this instruction, the participant is informed that colored labels indicate areas where the computer has identified play activities and that different colors correspond to different activities. The participant is told that colored labels are more likely to be correct than incorrect. Grey annotations indicate areas where the computer identified that the toys experienced motion but could either not classify the play or determined that no play was occurring. The participant is informed that the toys are often in motion as the result of being bumped, reverberating from previous actions, or that the toys are picking up vibrations from other movements in the play space. However, there is a guarantee that no motion exists in the spaces between the grey labels. The participant is urged to leverage computer assistance, whenever possible, to aid the search process.

The participant is shown how to identify computer supplied annotations for his specific three play activities within the activity view and how to skip forward to these activities to hasten the search process. He is also asked to double click on various annotation labels to demonstrate how to skip the videos forward and adjust the videos to that location. Next, the participant is instructed to locate two of each activity via the fast searching method.

After demonstrating the ability to locate labels, the participant is then shown how to adjust annotations both in the temporal and categorical sense. Dragging on the boundaries of an annotation can adjust the duration of the annotation by altering the beginning and ending points of the annotation. To reclassify an annotation label (*e.g.,* from banging to shaking) the participant simply selects the new categorization from the properties pane. In addition, the participant is also shown how to delete incorrect labels. If the participant encounters a computer supplied annotation that is incorrect, he is instructed to delete it. However, he is instructed to only check the correctness of labels concerning his three play activities. Labels can be deleted by selecting the annotation and hitting the delete key or by selecting delete from a context menu. At this time, the participant is asked to locate two labels and change their categorizations, adjust the duration of two annotations, and delete an annotation.

The remaining task to teach the participant is creating new annotations. A participant

needs to create annotations during the NONE condition and in other conditions if he encounters play activities that the computer failed to annotate. Annotations can be created in either the toy view or the activity view portion of the interface. The participant clicks in the appropriate track and drags forward in time until the activity ceases. He then right clicks and selects the appropriate categorization for the activity. Corrections to the categorization can be made in the properties pane if needed.

After a participant practices creating labels, he is then shown how to add the activity annotation to the log. Annotations are added (or removed) from the log by right clicking on the annotation label and selecting the appropriate action from a context menu. Annotations are logged after being inspected and adjusted to the appropriate duration. The participant is instructed to log both correct annotations provided by the computer, computer annotations that he adjusted, as well as any annotation he creates. Logging the activities helps determine which computer supplied annotation the participant viewed and which ones he did not. Prior to the conclusion of training the participant is asked to find an example of each of the three play activities and add them to the log (as well as remove one annotation from the log). The participant is instructed to use the interface and ask questions. The training session is complete when the participant states that he is comfortable with the interface.

7.3.2.3 Experimental Conditions

Each participant receives all four levels of recognition quality in permuted order as dictated by the partial Latin square. At the start of the each condition, the experimenter loads the appropriate play session and applies the appropriate level of automatic recognition support. The participant is then seated at the computer, given time to adjust the video components of the interface and ask any questions they may have about the task. The participant is told that he will have fifteen minutes to complete his task and that each play session is over forty–five minutes in length. The participant is told that he must balance the number of instances he views and the level of detail he uses to correct annotation durations in order to complete the task. The participants are also told that it is possible to reach the end of the video if he leverages the information provided by the computer. If the participant has

no further questions, he is told that he will receive verbal notification when he has a minute remaining to complete the task and to begin when ready. A screen capturing program is started by the experimenter just prior to the start of the search process.

At the end of the fifteen minutes, the experimenter saves the log and label annotations. The experimenter then starts a computerized version of the NASA–TLX survey [3]. When the participant completes the NASA–TLX, he is then administered a paper survey ascertaining task search strategy, task difficulty, and his satisfaction with the quality of the annotations provided by the computer (see Appendix E for the survey details). Once the participant has completed the survey, the experimenter scans the document and asks any clarifying questions, if needed. After the survey process is complete, the participant begins his next condition, repeating these steps.

7.3.2.4 Post–Conditions

After all four conditions and the associated surveys have been completed a post–experiment survey is administered. The participant ranks which condition he likes best, which condition provided the most useful annotation support, and which condition was easiest. The survey also compares conditions for similarity, and asks several questions about which aspects of automatic annotations (*e.g.,* annotation duration and categorization) are most important. The survey also investigates the impact of the software interface design on task performance. The participant is asked clarifying questions after the completion of this survey.

7.3.3 Participants and Compensation

Twenty participants, sixteen males and four females, were recruited from the Atlanta Metropolitan Area as well as from the student population of the Georgia Institute of Technology. Each participant is recruited for a single, two and a half hour session and receives $20 for his involvement. As a recruitment criteria, participants must be able to use a traditional mouse, understand verbal instructions, be able to distinguish between different colors displayed on a computer monitor, and have basic computing skills. This population can be considered similar to that which is often hired to annotate data for university based psychology research studies.

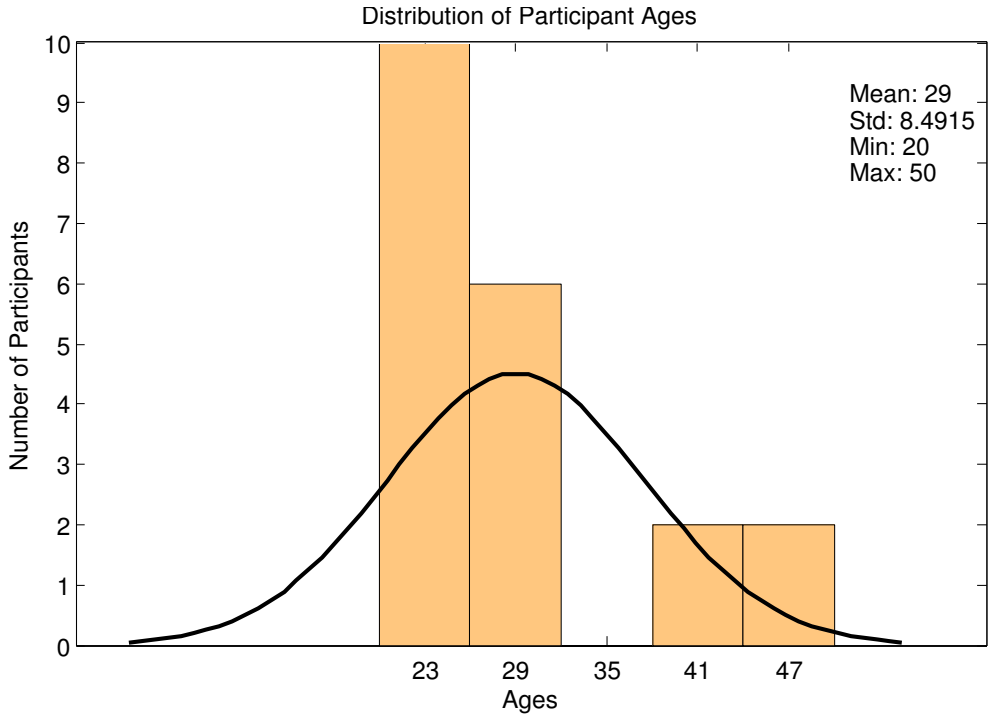


Figure 28: Histogram of participant's ages

Participants are between the ages of 20 – 50 years of age, with the mean age being 29. Figure 28 is a histogram of participant ages. All of the participants have at least a high school education. Eight of the participants were currently obtaining or held degrees in computer science. Only two participants reported having experience with pattern recognition or machine learning courses. Overall, participants report using a computer for 30 – 49 hours a week. Four participants have experience with video editing software (under 50 hours of use). No participants reported experience transcribing or annotating video.

None of the participants have been involved in previous clinical research involving young children. Only two of the participants are parents. However, all but three participants report experience watching young children play with toys.

7.4 Performance Measures

Several measurements are collected during the course of the study from each participant to help assess both quantitative and qualitative aspects of automatic play recognition support and the impact it has on identifying play activities. First, the time taken to identify all

instances of assembling Lego^{TM} Quatros, puppy jumping, and shaking any toy that exist in the play session is recorded. The expectation is that the level of annotation will directly influence the time required to annotate a play session. If participants do not finish annotating a play session in the given time, a second measure of performance is the percentage of the play session physically viewed by the participant. The furthest point reached in the play session is computed *post hoc* from screen captures collected during the session and is used to compute the percentage of the play session that is viewed. Third, regardless if the participants completely annotate the play session data, the resulting annotations are a combination of the computer generated labels and human supplied corrections. These annotations are recorded for each participant and accuracy metrics evaluating the quality of the annotations are computed. To help ascertain which labels are affected by the participant, a log file is created by each participant which details the annotations he has physically viewed, potentially adjusted, and verified for correctness. The log file is necessary as the computer may provide labels that do not require adjustment and would look identical to its state prior to the participant viewing it. The log allows the experimenter to distinguish between which annotations were viewed and required no corrections versus which annotations that the participant did not reach due to time constraints but are still included in the resulting annotation file.

In addition to metrics resulting from task performance, survey instruments are used to collect a quantitative measure of perceived effort and frustration caused by errors. These instruments include two post–condition surveys and an exit survey after the experiment is complete. In these surveys, the participants provide information via forced rankings, Likert scales, and short essay responses.

7.5 Analysis of Performance Metrics

Analysis of the data collected is presented according to task performance metrics and followed by survey data.

7.5.1 Play Identification Performance Metrics

During each condition, a participant is asked to log play events as well as remove mistakes made in the computer annotations. The resulting play annotation is a combination of computer generated annotations with human corrections. These annotations, which include the participants ability to identify play behaviors, can be evaluated in the same way that the statistical models of play recognition are evaluated.

The F_1 score is the harmonic mean of the positive prediction value and the true positive rate. A one–way repeated measures ANOVA was conducted to compare the F_1 scores of the participant's play annotations after completing the None, Motion–Only, Low, and the High annotation conditions. The means and standard deviations are presented in Table 23 as well as illustrated in Figure 29. There was a significant effect for annotation condition, Wilks' Lambda = .063, $F(3,17) = 84.01$, $p = .0005$, multivariate partial eta squared = .937.

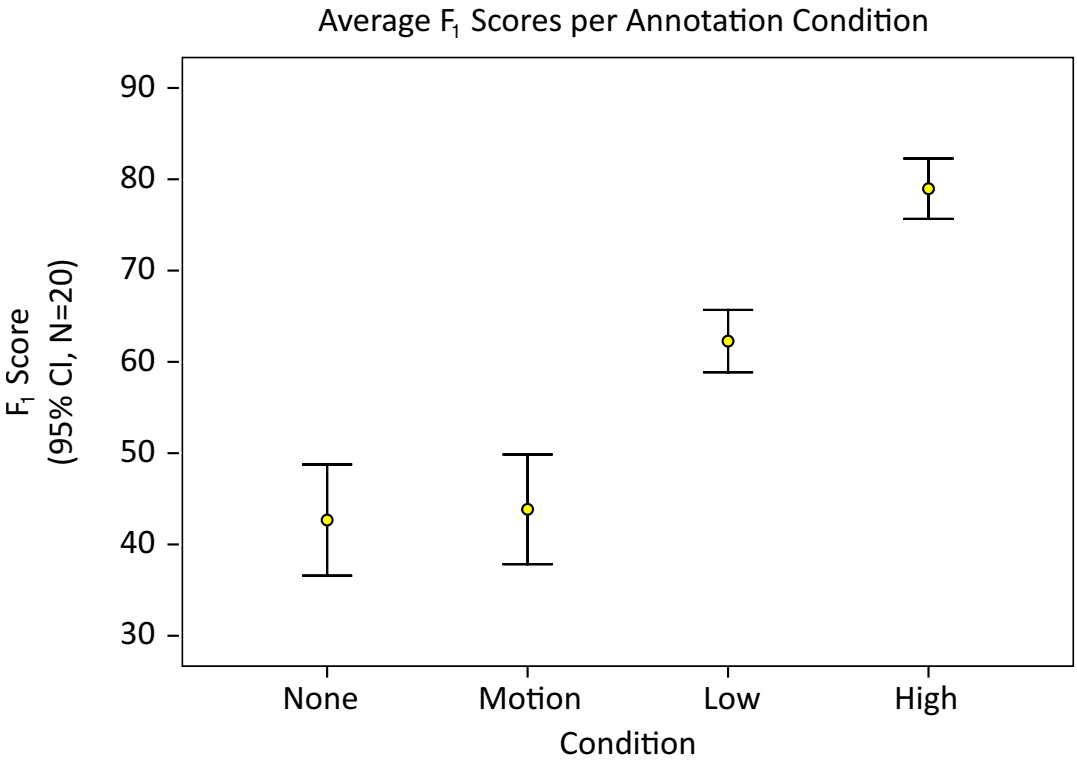


Figure 29: F_1 scores of participants' annotations grouped by condition.

The F_1 scores did not vary significantly between the None and the Motion–Only

Table 23: Descriptive statistics for F_1 scores of participants' annotations of play activities.

Condition	Participants (N)	Mean	Standard deviation
None	20	42.67%	13.01%
Motion–Only	20	43.85%	12.84%
Low	20	62.27%	7.30%
High	20	78.96%	7.06%

conditions. However, there is a significant difference between the Low and High conditions ($p = .000$); between Motion–Only and High ($p = .000$); between Motion–Only and Low ($p = .000$); between None and High ($p = .000$) as well as between None and Low ($p = .000$);

The results indicate that overall participant performance does not significantly change between the baseline (None) condition, which provides the user with no annotations, and the Motion–Only condition, which provides the user with annotations indicating toy motion. While the generalized motion labels help participants skip over areas of inactivity, the labels were too frequent and nondescript to help significantly increase performance. Of important note was that the motion labels often included non–play motions, such as toys being kicked and bumped. There were 12 participants that made comments regarding the slight benefits provided by the motion labels over the None condition. Participant g4p3 states, *"... I think I actually made it further into this video [*None *] and identified more activities than in the last [condition* Motion–Only *]. I think the grey labels in the last [condition] helped me a little but were not that descriptive and ended up having me waste my time and distract me so I stopped using them."* Similarly, Participant g5p3 states, *"... Grey boxes are only slightly more useful than not having anything especially when she [the child] is all wound up [jumping around causes motion in the toys when they are not being used]."*

The conditions with annotations provided by the computer significantly increase performance as the annotation condition increases in quality. This result is not surprising as the High condition was designed by modifying the different error types present in the Low condition until there was a significant increase in the F_1 measure. The average F_1 scores of the participants is higher, pairwise, than the original F_1 scores of the High and Low

conditions – hinting at a potential gain to keep humans in the play recognition loop while using the *Child'sPlay* system, even as future recognition abilities increase towards the HIGH condition (and potentially beyond). The system, as it stands, is significantly better than current best practices and can only improve as technology allows the system to move from LOW to HIGH.

7.5.2 Percentage of Video Reviewed

Unexpectedly, fourteen of the twenty participants did not reach the end of the play sessions during the HIGH and LOW conditions. The percentage of video reviewed by the participants is calculated using the screen captures recorded during the participants sessions. The starting position of the play–head is marked and the furthest position reached in the time line is considered the ending point. The differences between these two positions is calculated and divided by the length of the video.

Condition did not have a significant effect on the percentage of video viewed by the participant. I feel this result directly relates to the search strategy utilized by the participants. During training, each participant is shown how to maximally use the provided annotations to search through the play session. Participants must demonstrate the ability to search through the training video in a similar fashion before proceeding to the first experimental condition. However, only five of the twenty participants adopted the training search strategy in both the HIGH and LOW condition. One participant switched to the training strategy for his last condition (accounting for the six participants which viewed all of the play session video).

When questioned about the search strategy used, eight participants stated that they wanted the annotations that they provided to be as accurate as possible. For example, Participant g4p2 states, *"[if] I didn't finish the video, at least what I did finish was solid."* Participant g1p4 states, *"Given that the video was much longer than the time available, I didn't focus on getting through it, more on trying not to miss any play activities and trying to be accurate in the start and stop times."* Of the participants that watched large portions of the video in real–time, three participants stated that they chose to watch the video in

real–time (and near real–time) as it seemed most natural or was easiest. Participant g1p1 states, *"it just seemed the method that was most familiar to me."* One participant watched the video in near real–time due to a lack of trust in the recognition Participant g2p2 states, *"I didn't fully trust the labels so I wanted to view most of the video myself, and that seemed the likely best way to do it quickly and accurately."*

The degree of trust in computer algorithms has an impact on the participants' search process at some level. As part of the background survey all participants are asked for an opinion of the following statement: "If a computer uses an algorithm to provide me with information, I believe it to be correct. Figure 30 shows the distribution of responses. Four of the participants disagreed with this statement, five were neutral, and eleven agreed.

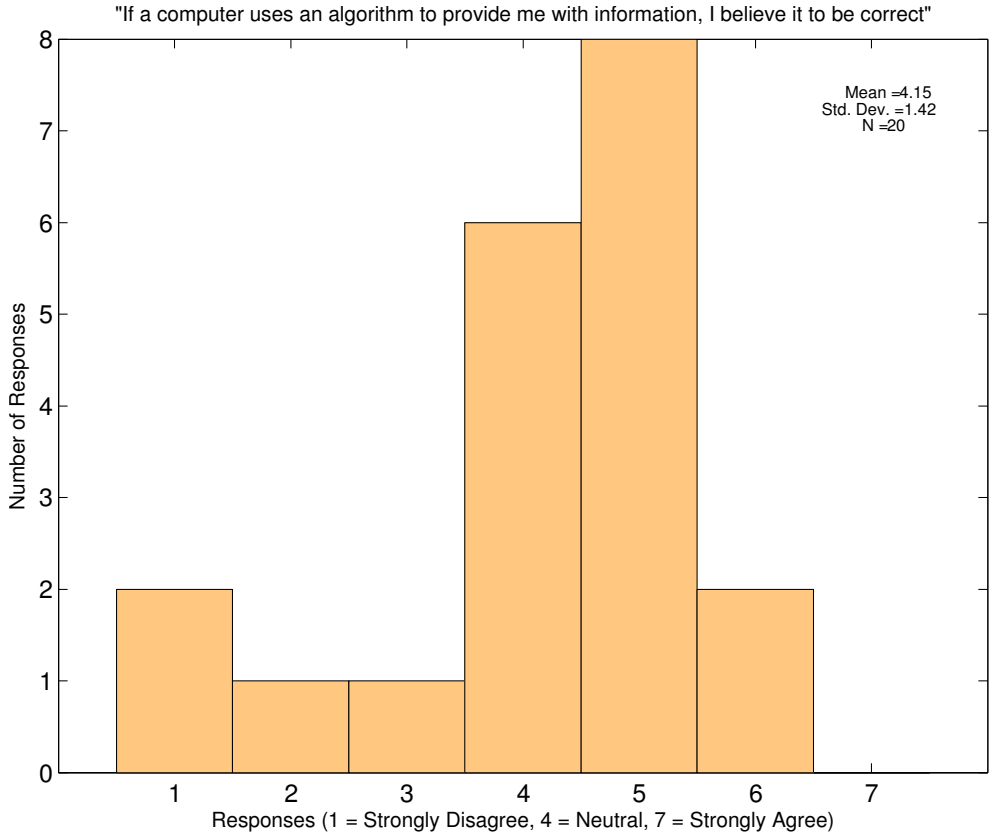


Figure 30: Histogram of responses to background questions 13: "If a computer uses an algorithm to provide me with information, I believe it to be correct."

This distribution of responses to the belief in computer supplied information may suggest that many participants wanted to maximize the accuracy of the labels they themselves

created to augment the computer labels, which they mostly trusted to be correct. In explaining his strategy, Participant g5p1 states, *"I decided to aim for finding as many [missing play activities] as I could rather than checking as many [computer provided labels] as I could. I was going for accuracy over speed. I was not ignoring the labels, but augmenting the existing labels and checking labels as I encountered them just to make sure they were OK."* No other participants expressly stated that they were specifically augmenting the existing labels. However, the search strategies of eight other participants were very similar to the strategy used by participant g5p1.

It should also be noted that the trust in the computer's accuracy did not change significantly as a result of the participants annotating four play sessions. A paired T-test was conducted on the pre–experiment beliefs and post-experiment beliefs and did not reveal a significant difference between the responses. Figure 31 compares the pre–experiment and post–experiment distributions.

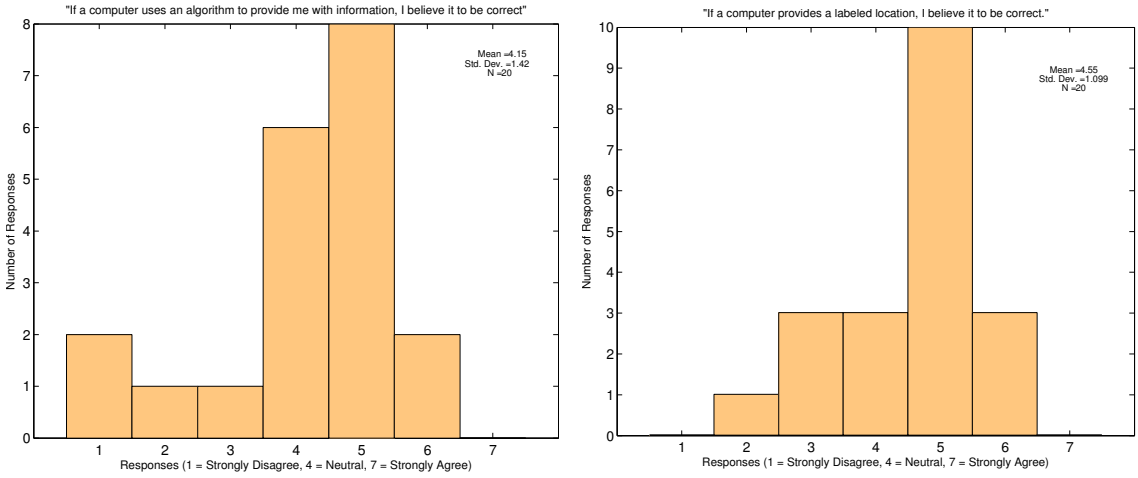


Figure 31: Comparison between pre–experiment and post–experiment belief in computer accuracies (from left to right)

7.5.3 Number of Logged Play Activities

In light of the search strategies discussed above, it is also important to investigate the effect that annotation condition has on the percentage of play activities logged. A one–way repeated measures ANOVA was conducted to compare the percentage of proper play instances logged by the participant after completing the None, Motion–Only, Low, and

the HIGH annotation conditions. The means and standard deviations are presented in Table 24 as well as illustrated in Figure 32. There was a significant effect for annotation condition, Wilks' Lambda = 0.430, $F(3,17) = 7.52$, $p = .002$, multivariate partial eta squared = 0.570.

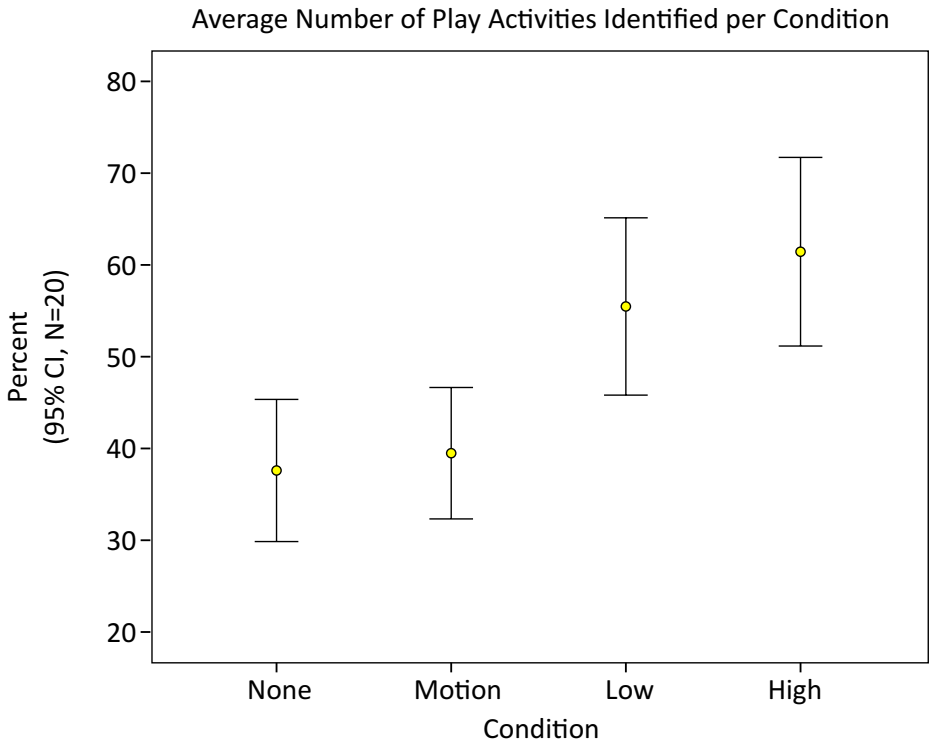


Figure 32: Percentage of play activities logged by participants over each condition.

Table 24: Descriptive statistics for the percentage of play instances logged

Condition	Participants (N)	Mean	Standard deviation
NONE	20	37.60%	16.55%
MOTION–ONLY	20	39.48%	15.30%
LOW	20	55.47%	20.64%
HIGH	20	61.44%	21.96%

The percentage of items logged did not vary significantly between the NONE or MOTION–ONLY conditions nor between the LOW and HIGH conditions. However, there is a difference between NONE and LOW ($p = .008$); NONE and HIGH ($p = .002$); MOTION–ONLY and LOW ($p = .012$); as well as MOTION–ONLY and HIGH ($p = .002$).

In terms of the number of events found, there is a significant difference between the

percentage of events found in conditions that provided activity specific annotation when compared to those that did not. This result helps justify systems such as *Child'sPlay* that use statistical models to help identify activity versus those that might rely solely on simple motion indicators and assume that a human can filter through the results to make intelligent annotations.

7.6 Analysis of Survey Data

In addition to performance metrics, it is also important to factor in the participant's perceptions. Surveys were conducted after a participant completed annotating each play session. Each participant completed a total of four Post–Condition surveys. Unless otherwise stated, responses are collected from a seven point Likert scale with a score of 1 = *Strongly Disagree* and a score of 7 = *Strongly Agree*. The next few sections, discuss these results.

7.6.1 Satisfaction with Annotation Support Provided by the Computer

A one–way repeated measures ANOVA was conducted to compare Likert scale scores on the participant's satisfaction with the number of labels provided by the computer after completing the None, Motion–Only, Low, and the High annotation conditions. The means and standard deviations of the participants' responses to the question, *"I am satisfied with the number of labels provided by the computer."* are presented in Table 25 as well as illustrated in Figure 33. There was a significant effect for annotation condition, Wilks' Lambda = .43, $F(3,17) = 7.39$, $p = .002$, multivariate partial eta squared = .95.

Table 25: Descriptive statistics for responses regarding the satisfaction with the number of labels provided by the computer

Condition	Participants (N)	Mean	Standard deviation
None	20	2.85	1.90
Motion–Only	20	3.60	2.04
Low	20	5.25	1.16
High	20	5.30	1.22

Satisfaction with the number of labels did not vary significantly between the None or Motion–Only conditions nor between the Low and High conditions. However, there is a difference between None and Low ($p = .001$); None and High ($p = .001$); Motion–Only

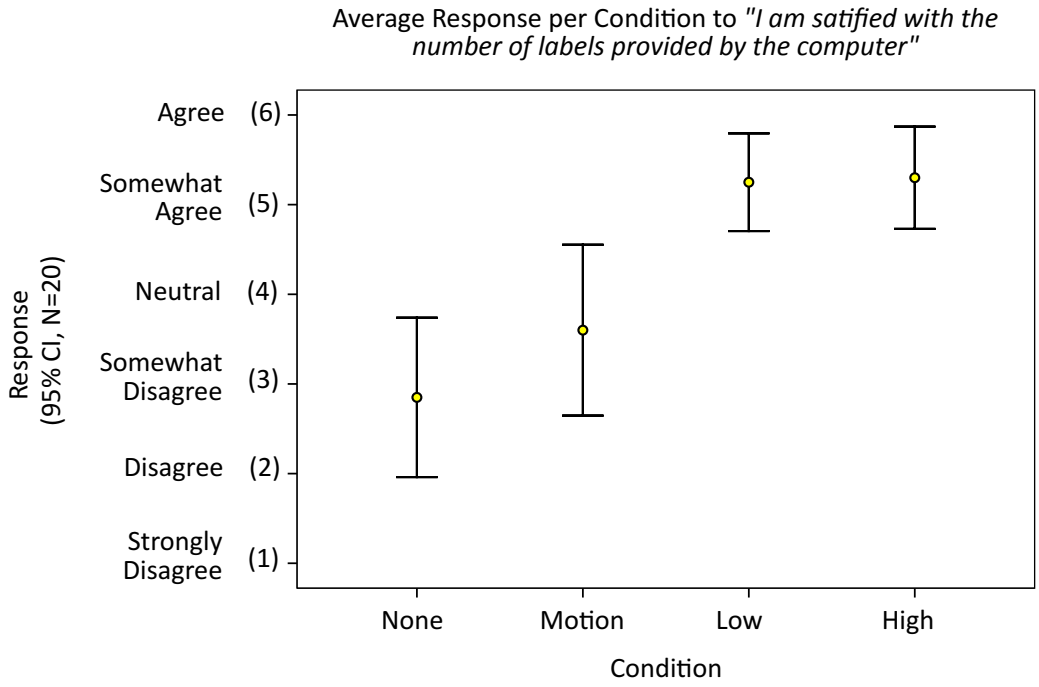


Figure 33: Average Likert scale response to "I am satisfied with the number of labels provided by the computer." grouped by condition.

and Low ($p = .017$); as well as Motion–Only and High ($p = .012$)).

These results indicate that, in terms of number of labels provided, participants viewed the grey Motion–Only labels as equivalent with no annotations and are, in general, dissatisfied by the number of labels provided. Participants were more satisfied with the number of detailed colored annotations in the Low and High conditions over just the grey motion labels alone. Participant g1p2 states, *"The grey labels were pretty useless, it seemed that they were triggered even the kid walked next to the toy."* While detailed color annotations are more satisfying than generalized motion labels, there is not a significant change in satisfaction between the Low and High quality annotations.

7.6.1.1 Searching in the Presence of Annotations for Multiple Activities

When interpreting the performance metrics above, it is also important to have an understanding if the participant felt overwhelmed by aspects of the task. In particular, the automatic recognition was not specifically tailored for the three play tasks that the participants were searching. The presence of other types of play causes more annotations to be

present and increases the opportunity for play events to be misclassified.

A one–way repeated measures ANOVA was conducted to compare scores on the participant's ability to ignore annotations not related to the primary search task after completing the None, Motion–Only, Low, and the High annotation conditions. The means and standard deviations of the participants' responses to the question, *"I was not distracted by labels not related to my task."* are presented in Table 26. There was a significant effect for annotation condition, Wilks' Lambda = .56, $F(3,17) = 4.53$, $p = .016$, multivariate partial eta squared = .44.

Table 26: Descriptive statistics for responses regarding the ability to ignore annotations not related to the primary search task

Condition	Participants (N)	Mean	Standard deviation
None	20	6.50	0.76
Motion–Only	20	5.30	1.90
Low	20	5.45	1.47
High	20	5.60	1.43

Ability to ignore unrelated annotations did not vary significantly between the Low or High conditions nor between the Low, Motion–Only, and High conditions. However, there is a difference between None and Low ($p = .024$) as well as None and Motion–Only ($p = .050$).

During the search tasks, participants are asked to locate three different types of play among thirteen total activities. These results indicate that participants can search for three different play activities without being distracted by information pertaining to the ten other types of play. Furthermore, this result is independent of the level of annotation provided by the computer.

7.6.2 Searching in the Presence of Inaccurate Annotations

In addition to gauging the impact of unrelated activities, it is also important to understand the impact that errors have on the search process. A one–way repeated measures ANOVA was conducted to compare scores on the participants' ability to perform the search task in the presence of inaccurate annotations after completing the None, Motion–Only, Low,

and the HIGH annotation conditions. The means and standard deviations of the participants' responses to the question, *"Erroneous labels did not prevent me from completing my task."* are presented in Table **??** as well as illustrated in Figure 34. There was a significant effect for annotation condition, Wilks' Lambda = .35, $F(3,17) = 10.46$, $p = .0005$, multivariate partial eta squared = .65.

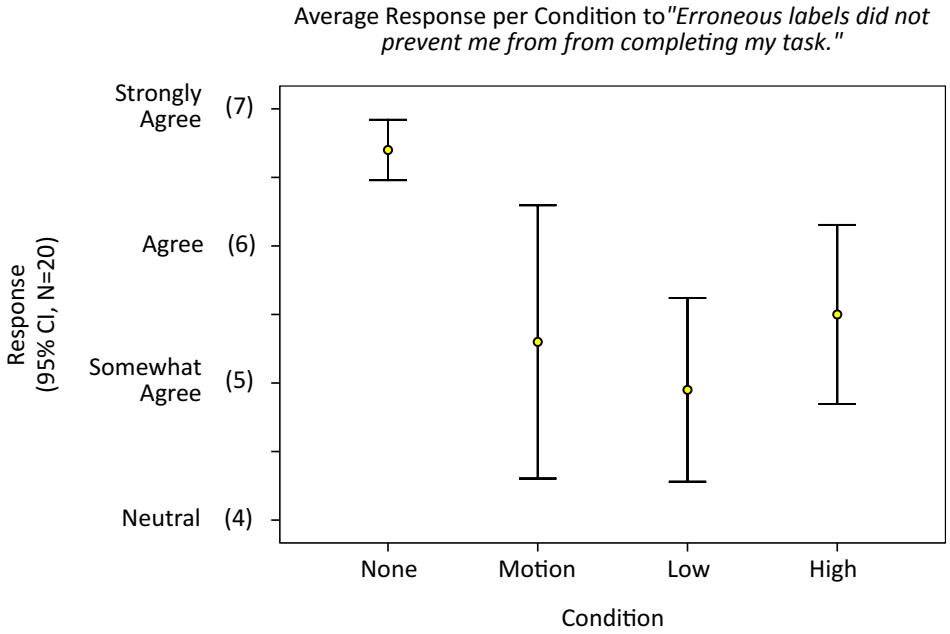


Figure 34: Average Likert scale response to "Erroneous labels did not prevent me from completing my task." grouped by condition.

Table 27: Descriptive statistics for responses reguarding the ability to perform the search task in the presence of inaccurate annotations

Condition	Participants (N)	Mean	Standard deviation
NONE	20	6.70	0.47
MOTION–ONLY	20	5.30	2.13
LOW	20	4.95	1.43
HIGH	20	5.50	1.40

The performance on search in the presence of inaccurate annotations did not vary significantly between the MOTION–ONLY, LOW and HIGH conditions. However, there is a difference between NONE and LOW ($p = .000$); as well as NONE and HIGH ($p = .008$). Recall that in the NONE condition, there are no annotations provided by the computer.

Inaccurate labels did not prevent the participants from identifying play activities. Though

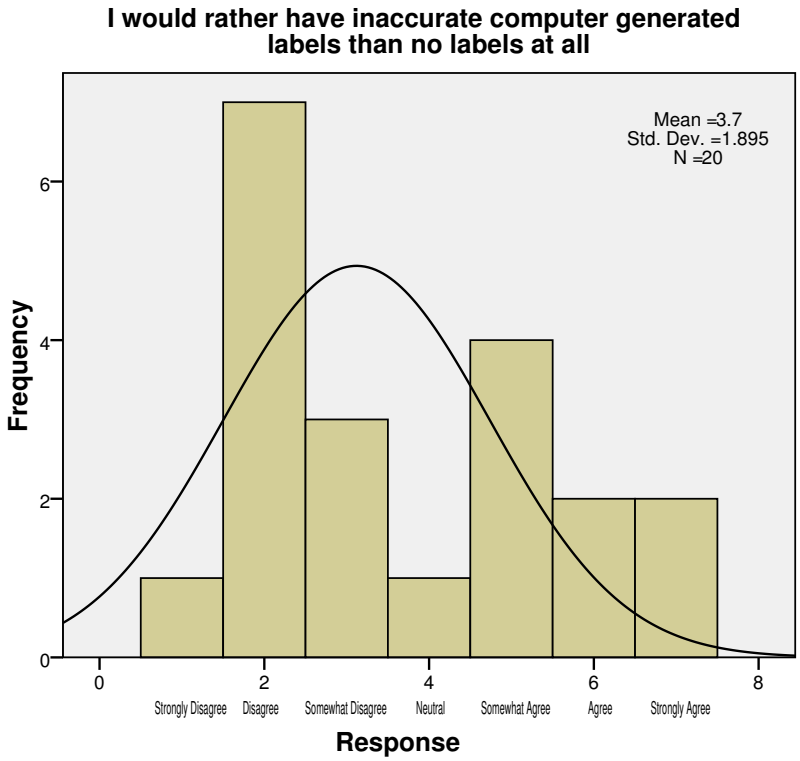
the difference between the baseline condition and the higher quality annotations levels suggests that there was more opportunity for erroneous labels to be encountered versus none at all. A lack of distinction between the MOTION–ONLY and NONE annotations here may indicate that the grey, MOTION–ONLY labels were not necessarily perceived as labels that were "correct" or "incorrect." Qualitative data supports two explanations. First, that the MOTION–ONLY annotations were merely a way to skip sections of inactivity within the play session. Participant g3p2 states, *"The labels at the top [grey] were useful for me when there were long breaks in toy movement and I would just skip over the dead space."* Second, the participants ignored the generalized motion annotations. Participant g6p2 states, *"I used grey to skip ahead but for the most part ignored them."* Participant g3p1 states, *"The grey labels seemed to be worthless."*

After completing all conditions, participants were again asked questions pertaining to the impact of errors during an exit survey. Figure 35 illustrates the distribution of responses to the question: *I would rather have inaccurate, computer generated labels than no labels at all.* Eleven participants disagreed with this statement while 9 agreed and one remained neutral.

The median response, 3.00, indicates that overall, participants, somewhat disagree with this statement. However, in the context of this response, I believe that participants interpreted "inaccurate labels" to mean the grey generalized motion labels associated with the MOTION–ONLY condition rather than classification errors within the conditions with colored annotations. This interpretation aligns with many of the negative comments directed towards the MOTION–ONLY and NONE conditions as well as positive comments directed towards the LOW and HIGH conditions. More evidence to support this interpretation is also provided by the response towards questions investigating false positive errors.

Figure 36 illustrates the distribution of responses to the question: *It is easier to ignore extraneous labels than it is to search video for missing labels.* Sixteen of the twenty participants agreed with this statement while two disagreed, and two remained neutral.

The median response, 6.00, indicates that participants, overall agreed with the above statement. This result suggests that participants would rather delete false positives than



N	20
Mean	3.70
Std. Error of Mean	.424
Median	3.00
Mode	2
Std. Deviation	1.895
Variance	3.589
Minimum	1
Maximum	7
Percentiles	
25	2.00
50	3.00
75	5.00

	Frequency	Percent
Strongly Disagree	1	5.0
Disagree	7	35.0
Somewhat Disagree	3	15.0
Neutral	1	5.0
Somewhat Agree	4	20.0
Agree	2	10.0
Strongly Agree	2	10.0
Total	20	100.0

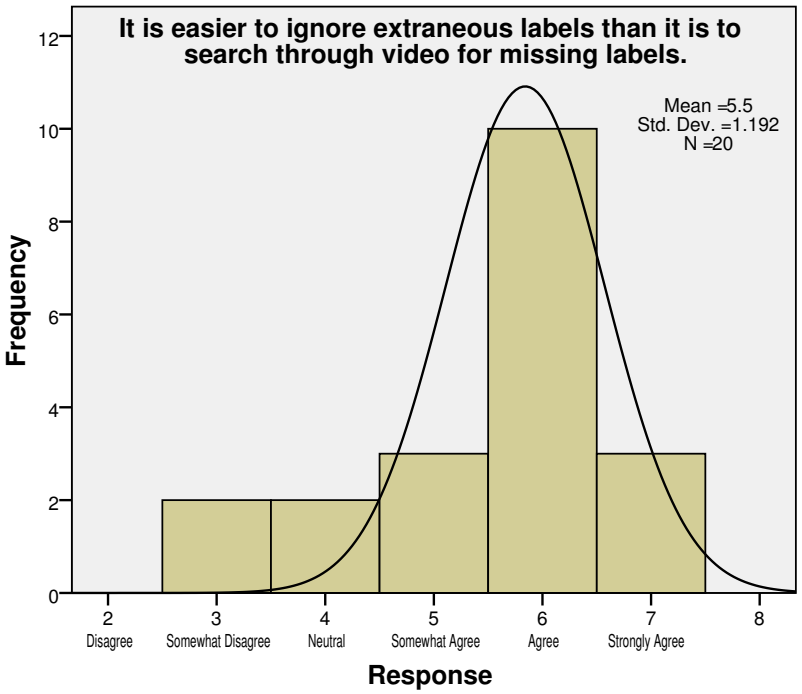
Figure 35: Histogram of responses pertaining to generalized annotations versus no annotations

receive no annotation support from the computer. Participant g5p3 states, *"If one is trying to produce an essay with few mistakes, it saves a lot of time if there is already a rough draft. Just correcting mistakes is easier than having to do the work yourself and then correct the mistakes."*

When interpreting both the response to this question and the previous question, one must be more specific about the types of errors the participants are willing to tolerate to avoid contradiction. In particular, the false positives must provide more specific information than simply informing the user that motion occurs, and occurs less frequently, than the general motion labels.

7.6.3 Computer Generated Annotations are Useful to the Search Process

Another way to confirm the interpretation about errors presented in the last section is to determine if participants found annotations that were useful during each of the conditions.



N	20
Mean	5.50
Std. Error of Mean	.267
Median	6.00
Mode	6
Std. Deviation	1.192
Variance	1.421
Minimum	3
Maximum	7
Percentiles	
25	5.00
50	6.00
75	6.00

	Frequency	Percent
Somewhat Disagree	2	10.0
Neutral	2	10.0
Somewhat Agree	3	15.0
Agree	10	50.0
Strongly Agree	3	15.0
Total	20	100.0

Figure 36: Histogram of responses pertaining to preference of insertion errors to deletion errors

A one–way repeated measures ANOVA was conducted to compare scores on the participant finding computer generated labels that are useful to his process after completing the None, Motion–Only, Low, and the High annotation conditions. The means and standard deviations of the participants' responses to the question, *"I found computer generated labels that were useful to my search."* are presented in Table 28. There was a significant effect for annotation condition, Wilks' Lambda = .105, $F(3,17) = 48.19$, $p = .000$, multivariate partial eta squared = .89.

Table 28: Descriptive statistics for responses reguarding the presence of useful computer generated annotations

Condition	Participants (N)	Mean	Standard deviation
None	20	6.70	0.47
Motion–Only	20	5.30	2.13
Low	20	4.95	1.43
High	20	5.50	1.40

The perception of the presence of useful computer generated annotations did not vary significantly between the Motion–Only and Low conditions, the Motion–Only and

HIGH conditions, nor the LOW and HIGH conditions. However, there is a difference between MOTION–ONLY and HIGH ($p = .009$) as well as NONE and all other conditions (MOTION–ONLY $p = .001$, LOW $p = .000$, and HIGH $p = .000$).

In all levels of computer provided annotation, the participants were able to find annotations that assisted them in the search process. Of particular interest is the fact that the ability to find useful labels did not vary significantly between the LOW and HIGH conditions. While there were also no differences between the LOW and MOTION–ONLY conditions, there is a perceived benefit over just motion alone if the annotations are of sufficiently high quality (approaching quality found in the HIGH condition).

7.6.3.1 Confidence in Identifying All Instances of Play That Exist

Results from the previous sections suggest that participants prefer higher quality annotations over generalized motion labels or nothing at all. Next, it is interesting to investigate the impact that different quality annotations have on participants' confidence.

A one–way repeated measures ANOVA was conducted to compare scores on the participant's confidence in logging all existing play activities after completing the NONE, MOTION–ONLY, LOW, and the HIGH annotation conditions (survey question 12). The means and standard deviations of the participants' responses to the question, *"I am confident that I logged all instances of my play activities that exist in the video."* are presented in Table 29. There was a significant effect for annotation condition, Wilks' Lambda = .513, $F(3,17) = 5.39$, $p = .009$, multivariate partial eta squared = .49.

Table 29: Descriptive statistics for responses regarding the confidence of logging all existing play activities

Condition	Participants (N)	Mean	Standard deviation
NONE	20	2.70	1.84
MOTION–ONLY	20	2.65	1.73
LOW	20	3.75	1.74
HIGH	20	3.40	1.88

Confidence in identifying all existing play did not vary significantly between the LOW and HIGH conditions nor between the MOTION–ONLY and NONE conditions. However, there is a difference between the MOTION–ONLY and LOW condition ($p = .012$) .

These results indicate that the increase in annotation quality does not necessarily increase the confidence in identifying all instances of play. However, providing more play specific detail in relation to general motion information can help improve confidence. Overall, confidence is low across all conditions. This is partially related to both trust in the computer recognition and the timed nature of the task. Only six of the twenty participants were able to complete annotating a play session (in the LOW and HIGH condition). While more than half of the participants (12) expressed confidence in identifying all the behaviors in the percentage of the session they viewed, they acknowledged that they had not completed enough of the session to state with confidence that they had logged all instances of an event. Participant g1p3 states, *"I am not overly confident. It is a long video."*

Of the six that finished the video, all but one participant reviewed the video to identify play activities in the grey areas. Participant g4p3 states, *"[I have less confidence] because I only listened to the computer and it may have missed something."*

Participant g5p2 is the only participant that specifically stated that he lacked confidence in the recognition because he did not understand how the recognition worked beyond the simple explanation provided in training. Participant g5p2 states, *"I could have missed a shake I was scrolling really fast. I am 90% confident in [finding] all assembling I can imagine the sensors are pretty accurate with identifying that. I just don't know how the sensors can tell the difference from the puppy being shaken versus thrown across the room ... I don't know how confident I can be in the computer's ability to recognize [correctly] since this is all I looked at. I don't know if the computer missed something in the grey areas."*

The statement quoted above also suggest that the participants in this study may be more familiar with the capabilities of technology than the general population. This knowledge of technology may have made them increasingly more aware of the difficulty of the computation and the possibility that the underlying algorithms may not work as intended. Further study is needed to determine if the awareness of computation decreases trust in the automatic annotations. This issue will be discussed further in Chapter 9.

7.6.4 The Computer Reduces the Amount of Effort Required to Annotate

Given the lower scores in confidence for the computer generated annotations, it is important to investigate if participants feel the computer annotations reduce the effort required to annotate the video.

A one–way repeated measures ANOVA was conducted to compare scores on the participants' opinions that the computer reduces the amount of effort required to annotate play session after completing the NONE, MOTION–ONLY, LOW, and the HIGH annotation conditions. The means and standard deviations of the participants' responses to the question, *"Overall, the computer reduced the amount of effort required to annotate this video."* are presented in Table 30 as well as illustrated in Figure 37. There was a significant effect for annotation condition, Wilks' Lambda = .395, $F(3,17) = 8.66$, $p = .001$, multivariate partial eta squared = .61.

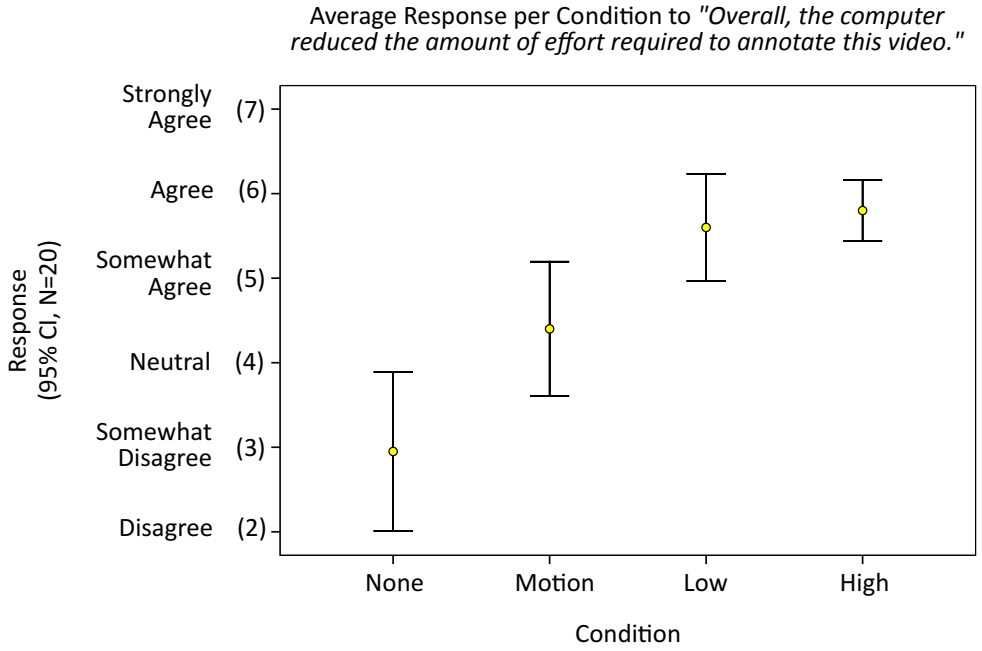


Figure 37: Average Likert scale response to "Overall, the computer reduced the amount of effort required to annotate this video." grouped by condition.

Opinions pertaining to required effort as a result of the level of computer provided support did not vary significantly between the LOW and HIGH conditions nor between the

Motion–Only and Low conditions. However, there is a difference between the Motion–Only and High condition ($p = .028$). There is also a difference between the baseline, None condition and all other conditions (Motion–Only $p = .024$, Low $p = .001$, and High $p = .000$).

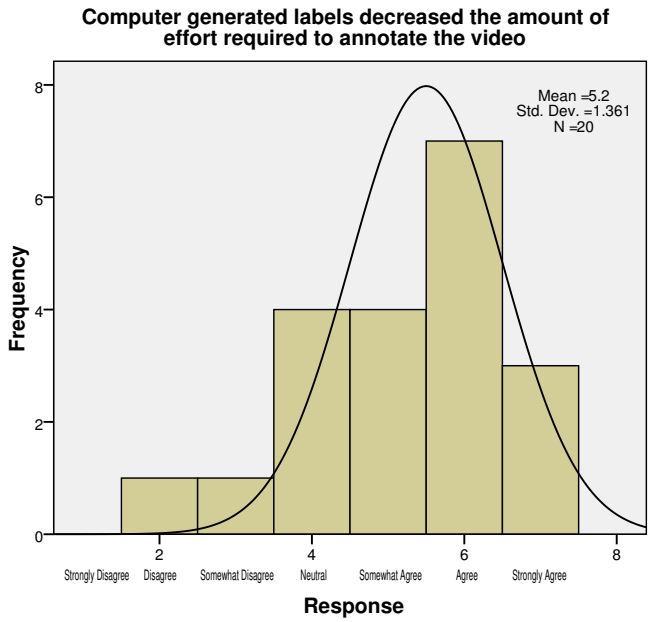
For all levels of annotation, the computer provided annotations significantly reduced the amount of perceived effort to annotate the data over not having any annotations provided. There is a significant decrease in effort again when comparing the effort required to annotate the data with general motion annotations compared to the highest quality annotations. In general, the grey labels allowed participants to ignore inactivity in the play data while the higher quality labels allowed participants to focus attention on specific types of play. It should be noted that there was not a significant decrease in effort between the lower quality annotations and the general motion labels. This lack of difference is attributed to the high insertion errors inherent in the Low condition (as described in Section 7.3.1.3). While the colored labels allow a participant to focus on specific play, there are several inaccurate instances that need to be discarded. These inaccuracies are not present in the High condition.

7.6.4.1 Least Effort and Overall Worklaod

Because the computer generated annotations are designed to reduce effort, the participants were asked, again, in the exit survey to both rank the conditions in terms of required effort as well as asked their overall opinion on effort. Figure 38 shows the distribution of response to the question *"Computer generated labels decreased the amount of effort required to annotate video."* Fourteen participants agreed with this statement, two disagreed, and four remained neutral.

Table 30: Descriptive statistics for responses reguarding the computer reducing the effort required to annotate play sessions

Condition	Participants (N)	Mean	Standard deviation
None	20	2.95	2.02
Motion–Only	20	4.40	1.70
Low	20	5.60	1.35
High	20	5.60	0.76



N	20
Mean	5.20
Std. Error of Mean	.304
Median	5.50
Mode	6
Std. Deviation	1.361
Variance	1.853
Minimum	2
Maximum	7
Percentiles	
25	4.00
50	5.50
75	6.00

	Frequency	Percent
Disagree	1	5.0
Somewhat Disagree	1	5.0
Neutral	4	20.0
Somewhat Agree	4	20.0
Agree	7	35.0
Strongly Agree	3	15.0
Total	20	100.0

Figure 38: Distribution of responses to "Computer generated labels decreased the amount of effort required to annotate video"

Of more interest are the rankings provided. Each participant ranked the conditions in order of required effort. The variance among the rank results is analyzed using a non–parametric alternative to the one–way repeated measures ANOVA, the Friedman Test. *Post hoc* analysis to determine the significant aspects is performed with the Wilocoxon Sign Rank Test which is the non–parametric alternative to the repeated measures paired–T test and has the advantage of comparing ranks instead of means [53].

Table 31: Descriptive statistics for the rankings of the condition that participants felt required the least effort (1 = least effort, 4 = most effort)

						Percentiles		
	N	Mean	Std. Deviation	Minimum	Maximum	25th	50th (Median)	75th
None	20	3.45	.887	1	4	3.00	4.00	4.00
Motion	20	2.85	.813	1	4	3.00	3.00	3.00
Low	20	2.00	.918	1	4	1.00	2.00	2.00
High	20	1.70	.979	1	4	1.00	1.00	2.00

The results of the Friedman Test indicated that there is a statistically significant difference in rankings of the condition which the participants felt required the least effort: $\chi^2(3, n = 20) = 22.98$, $p = .000$. Table 31 reports the descriptive statistics of the rankings per condition. Figure 39 illustrates the distribution of rankings across the conditions.

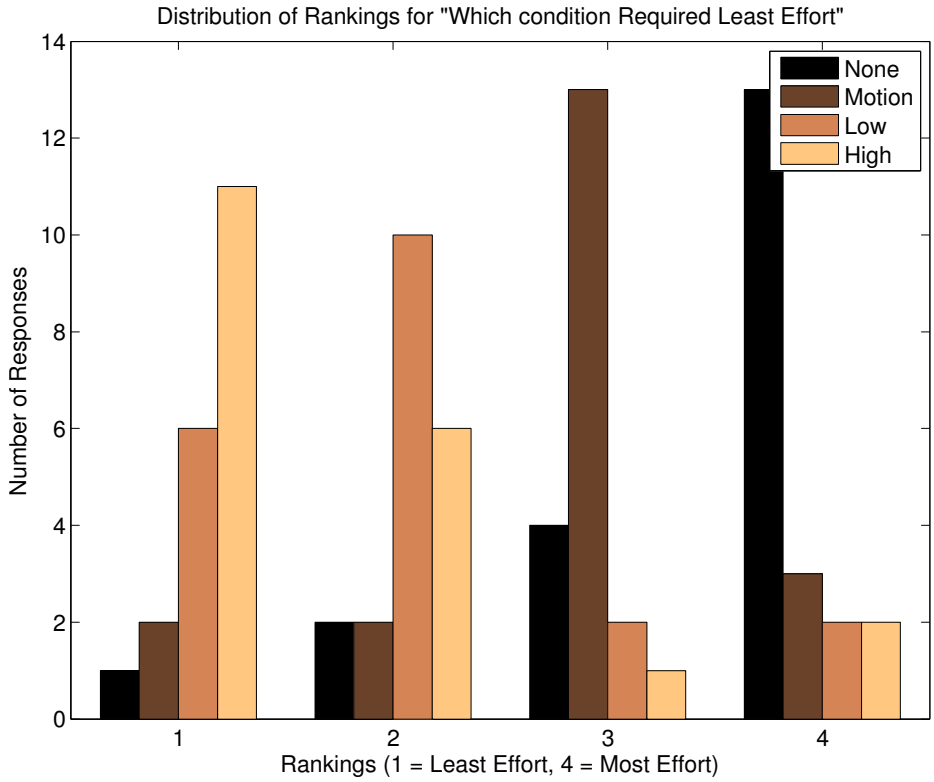


Figure 39: Distribution of rankings for the condition that participants felt required the least effort (1 = least effort, 4 = most effort)

A Wilcoxon Signed Rank Test revealed two statistically significant rank increases. First, there is an increase between the NONE and LOW condition, $z = -3.13$,p = .002, with a small effect size (r=.29). The median rank increased from condition NONE (Md =4.00) to LOW (Md = 2.00). Second, there is an increase between the NONE and HIGH condition, $z = -3.37$,p = .001, with a medium effect size (r=.31). The median rank increased from condition NONE (Md = 4.00) to HIGH (Md = 1.00).

From these results, it can be seen that the participants felt the HIGH and LOW conditions require less effort than the NONE condition. Again, while the median values differ, there is not a significant difference in effort between the NONE and MOTION–ONLY conditions, between the LOW and HIGH conditions, nor between the MOTION–ONLY and LOW conditions.

In addition to ranking conditions based of effort, the participants were also administered the NASA–TLX after each condition. The NASA–TLX is a tool used to measure perceived workload, and it was administered to measure differences in perceived effort, frustration, mental demand, performance, physical demand, and temporal demand. A one–way repeated measures ANOVA was conducted to compare responses on the NASA–TLX after completing the NONE, MOTION–ONLY, LOW, and the HIGH annotation conditions. There were no significant difference between the individual components. However, there was a significant effect for annotation on perceived overall workload, Wilks' Lambda = .540, $F(3,17) = 4.26$, $p = .023$, multivariate partial eta squared = .46.

Opinions pertaining to overall workload as a result of the level of computer provided support decreased significantly between all annotations conditions ($p = .000$), with MOTION–ONLY being the heaviest workload and HIGH being the lightest. The perceived workload, however, did not vary significantly between the NONE and MOTION–ONLY conditions. These results agree with the idea that as quality of annotation increases, the workload required by the participant decreases.

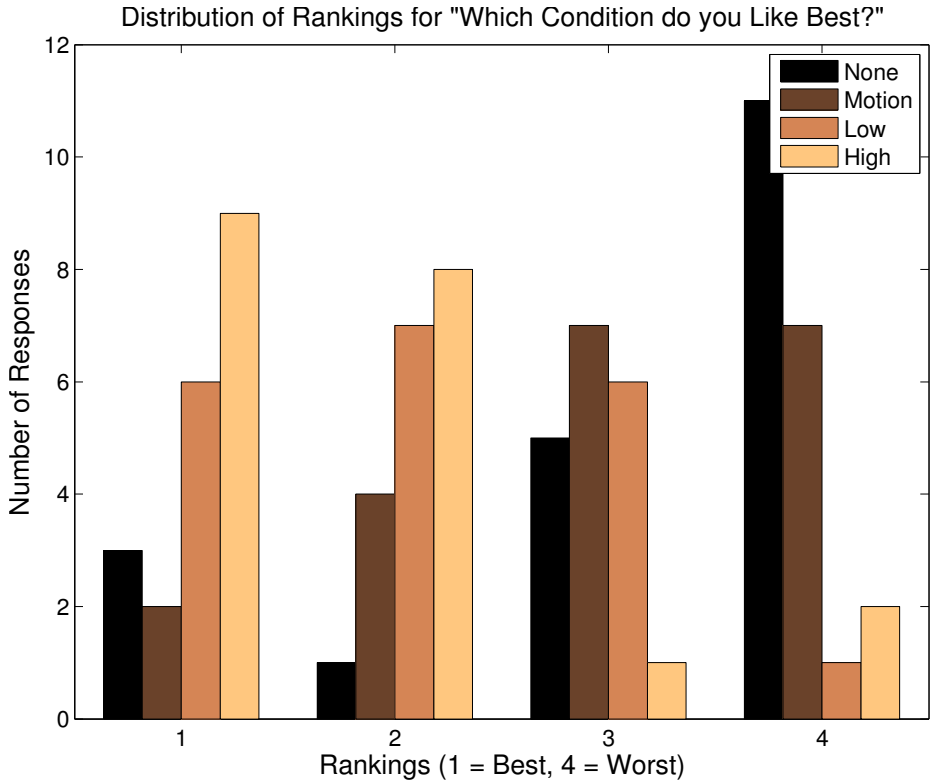


Figure 40: Distribution of rankings for the condition that participants felt was the best condition (1 = best, 4 = worst)

7.6.5 Best Condition Overall

In addition to ranking the quality of annotation in terms of effort, participants were also asked to rank the conditions in terms of the ones they like best.

Table 32: Descriptive statistics for the rankings of the conditions that participants liked best (1 = best, 4 = worst)

						Percentiles		
	N	Mean	Std. Deviation	Minimum	Maximum	25th	50th (Median)	75th
None	20	3.20	1.105	1	4	3.00	4.00	4.00
Motion	20	2.95	.999	1	4	2.00	3.00	4.00
Low	20	2.10	.912	1	4	1.00	2.00	3.00
High	20	1.80	.951	1	4	1.00	2.00	2.00

The results of the Friedman Test indicated that there is a statistically significant difference in rankings of which condition the participants liked best: $\chi^2(3, n = 20) = 16.21$, $p = .001$. Table 32 reports the descriptive statistics of the rankings per condition. Figure 40 illustrates the distribution of rankings across the conditions.

A Wilcoxon Signed Rank Test revealed a statistically significant rank increase between the NONE and HIGH condition, $z = -3.16$, $p = .002$, with a small effect size (r=.28). The median rank increased from condition NONE (Md = 4.00) to HIGH (Md = 2.00). There is also a statistically significant rank increase between the MOTION–ONLY and HIGH condition, $z = -2.61$, $p = .009$, with a small effect size (r=.26). The median rank increased from MOTION–ONLY (Md = 4.00) to condition HIGH (Md = 2.00).

These results indicate that the HIGH condition is liked better than both the baseline NONE and the MOTION–ONLY conditions. In other words, the presence of high quality annotations is liked better than the current best practices. There is not a significant difference between the rankings of the LOW and HIGH conditions, nor between the MOTION–ONLY and the LOW conditions. Again, as with previous questions, the lack of differentiation between the LOW and MOTION–ONLY conditions is attributed to the increased number of false positives between the LOW and HIGH conditions.

7.6.6 Most Useful Annotations

To help distinguish what participants like best about the conditions, participants were also asked to rank conditions in terms which condition provided the most useful annotations.

Table 33: Descriptive statistics for the rankings of the conditions in which the participants found the most useful annotations (1 = most useful, 4 = least useful)

						Percentiles		
	N	Mean	Std. Deviation	Minimum	Maximum	25th	50th (Median)	75th
None	20	3.55	.826	1	4	3.00	4.00	4.00
Motion	20	2.90	.718	1	4	3.00	3.00	3.00
Low	20	2.15	.988	1	4	1.25	2.00	2.75
High	20	1.40	.598	1	3	1.00	1.00	2.00

The results of the Friedman Test indicated that there is a statistically significant difference in rankings of the condition which the participants felt provided the most useful annotations: $\chi^2(3, n = 20) = 31.14$, $p = .000$. Table 33 reports the descriptive statistics of the rankings per condition. Figure 41 illustrates the distribution of rankings across the conditions.

A Wilcoxon Signed Rank Test revealed three statistically significant rank increases.

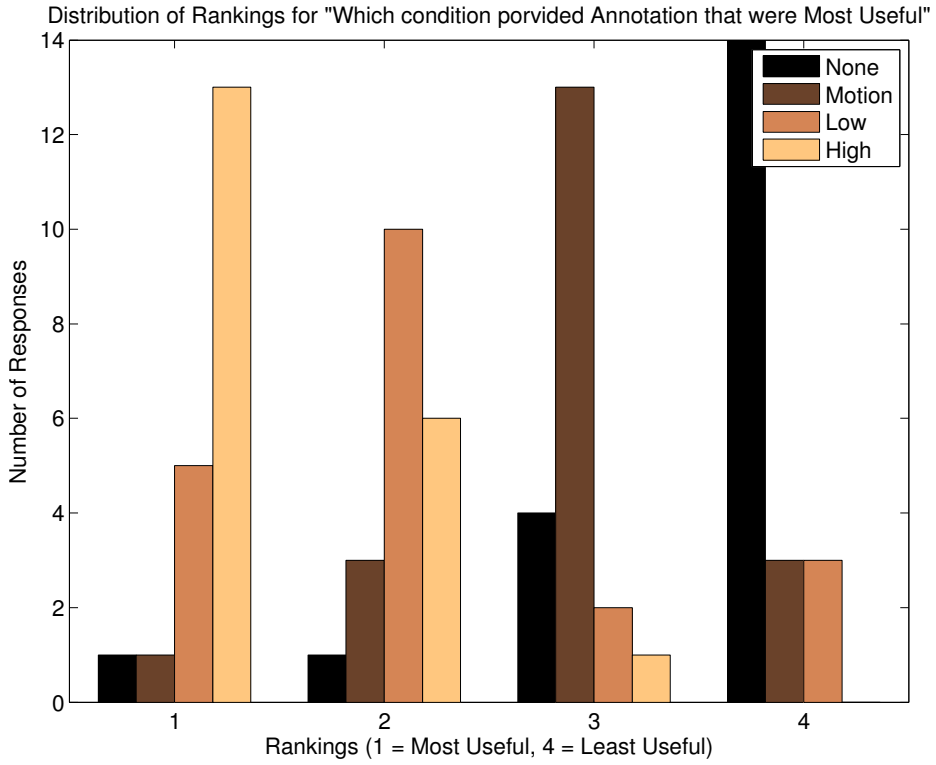


Figure 41: Distribution of rankings for the condition that participants felt provided the most useful annotations (1 = most useful, 4 = least useful)

First, there is an increase between the NONE and LOW condition, $z = -2.97$, $p = .003$, with a small effect size (r=.27). The median rank increased from condition NONE (Md = 4.00) to LOW (Md = 2.00). Second, there is an increase between the NONE and HIGH condition, $z = -3.89$, $p = .000$, with a medium effect size (r=.36). The median rank increased from condition NONE (Md = 4.00) to HIGH (Md = 1.00). Third, there is also a statistically significant rank increase between the MOTION–ONLY and HIGH condition, $z = -3.49$, $p = .000$, with a medium effect size (r=.31). The median rank increased from MOTION–ONLY (Md = 2.00) to condition HIGH (Md = 1.00).

These results indicate that both the HIGH and LOW conditions provide more useful annotations than the baseline NONE condition. Participants also found the HIGH conditions annotations to be more useful than annotations provided by the MOTION–ONLY condition. Although the median scores differ, there is not a significant difference in usefulness between the LOW and HIGH conditions, nor between the MOTION–ONLY and the LOW conditions. Not surprisingly, these results are similar to the rankings for the condition that participants

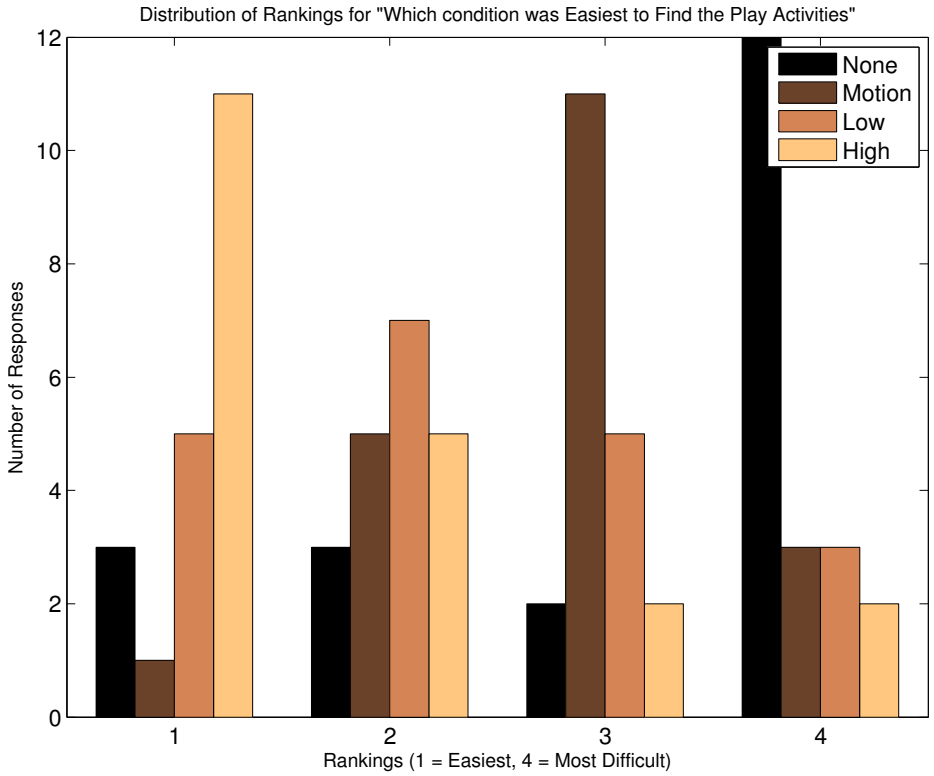


Figure 42: Distribution of rankings for the condition participants felt made it easiest to find play activities (1 = easiest, 4 = hardest)

liked the best.

7.6.6.1 Annotations that made it easiest to find play activities

While the distinction between best, most useful, and easiest are slight, participants were also asked to rank the conditions in terms of which condition made it easiest to find play activities.

Table 34: Descriptive statistics for the rankings of the condition which made it easiest to find play activities (1 = easiest, 4 = hardest)

						Percentiles		
	N	Mean	Std. Deviation	Minimum	Maximum	25th	50th (Median)	75th
None	20	3.15	1.182	1	4	2.00	4.00	4.00
Motion	20	2.80	.768	1	4	2.00	3.00	3.00
Low	20	2.30	1.031	1	4	1.25	2.00	3.00
High	20	1.75	1.020	1	4	1.00	1.00	2.00

The results of the Friedman Test indicated that there is a statistically significant difference in rankings of the condition which the participants felt provided the most useful

annotations: $\chi^2(3, n = 20) = 13.38$, $p = .004$. Table 34 reports the descriptive statistics of the rankings per condition. Figure 42 illustrates the distribution of rankings across the conditions.

While the nonparametric ANOVA is significant, after controlling for Type 1 errors, using a Bonferonni adjusted alpha value of .0083, the significance is not evident in the *post hoc* pairwise comparison tests. The variations in search strategies may account for the inability to differentiate the effect of annotation quality on ease of play identification. The variations may be a result of novice behavior – understanding expert use of the system is valuable and the future study of these users will be discussed further in Section 8.6.

7.7 Discussion and Future Implications for the Child'sPlay *System*

Overall, the participants liked the *PlayView* interface and the provided recognition support. The implications for the *Child'sPlay* system are two–fold in terms of the participant's performance and the participant's preferences. In terms of performance, the more specific annotations (HIGH and LOW) helped increase play retrieval over lesser quality annotations (MOTION–ONLY) or none at all which are the current standards. In terms of performance, there is a benefit to using the statistical models with an F_1 score = 48.0% to retrieve play activities over using naïve motion analysis on all sensor streams. The benefit of the statistical models will likely also increase as accuracy, in terms of the F_1 score, increases. These increases will be brought about as algorithmic or data sets advances allow the statistical models to achieve rates comparable to the HIGH condition, where the F_1 score = 70.0% or better.

In terms of participant preferences, the more specific the annotation quality, the better. In terms of developing the *Child'sPlay* system for future use, several factors revealed during this study are important to consider and investigate further:

1. **Confidence:** Users of the retrospective system must have confidence in the annotations provided by the system. The level of confidence in the information provided by the computer impacts the search strategy used. If the user does not have confidence in the annotations, they will be of little benefit to the user. While this is not a new

concept to the design of intelligent systems, it is important to reiterate. Trust in the computer supplied annotations comes from understanding how the computer can generate accurate annotations and understanding why it selected the annotation that it did. Belief in the labels provided by the computer either allows the user to "trust" the already existing annotations and look for other instances, or quickly review those instances provided by the computer, knowing that it retrieved a majority of the events of interest.

2. **Annotation Usage:** Users of the *Child'sPlay* system will likely use annotations in different ways. Annotations may be used merely as an interest–point indicators, alerting users when to pay attention while scanning through data. Annotations can also be used to help differentiate the difference between visually similar activities (similar to how participants used annotations to distinguish shaking from banging).

3. **Contextualize Error Types:** Different types of errors have different impacts depending on the quality of the annotation. If the annotation is too vague, consisting of elongated merge and substitution errors, participants may prefer no assistance at all. If the false positive errors are in the presence of more specific annotations, and the errors are also specific and concise, they are more tolerable. Search strategy again, can impact the preference between error types – especially with regards to boundary errors. Participants that watch the video in near–real time may prefer annotation creation, while those that jump ahead to only annotations prefer adjustment and deletion of errors.

4. **Hierarchical Annotations:** Participants search in a variety of ways and handle errors in different ways. A combination of the generalized motion annotations and very specific annotations can promote different types of investigation of the data and play event retrieval performance.

CHAPTER VIII

DISCUSSION AND FUTURE WORK

In this dissertation I have presented several augmented toy designs as well as explored a variety of supervised learning methods to facilitate both the automatic collection and automatic recognition of object play. The automatic collection and identification of object play data is challenging for a variety of reasons. This chapter will recapitulate the main challenges associated with recognizing object play, how specific design choices in the *Child'sPlay* system addresses many of these challenges, and discusses future directions for the technology explored in this dissertation.

8.1 Challenges in Object Play Recognition

A key aspect of recognizing different levels of object play sophistication is being able to automatically differentiate between play versus non–play activities. Distinguishing play from non–play is very difficult using limited motion sensing capabilities in combination with multiple augmented toys. Consider a child laying on the ground kicking an upside down plastic dome toy with her foot while she plays with two other toys in her hands. Using only data from the augmented toys, it could appear as if the three toys were being used together because all three toys are experiencing high motion. However, in reality, only two toys are engaged in object play while the plastic dome is simply reverberating from being kicked. In this case, being able to extract the object play from the motion caused by accidental bumps and fidgets is very challenging as it happens concurrently, and each toy is experiencing enough motion that both sophisticated and naïve motion filters would identify it as important. As discussed in sections Section 5.6 and Section 6.4.1 correlative features can help recognition systems distinguish between toys being used together versus toys that are used in an unrelated manner; however, it does not solve the problem. In addition, these non–play activities happen more frequently than expected as children often roll, flop, jump, and crawl around play spaces paying little attention to toys that they topple or bump in

the process.

The extraction and distinction of play activities from non–play activities is further complicated by the fact that there is not a single, salient feature of play that can be used to distinguish toys experiencing playful motions from those that are not. The salient aspects of play that can be used to distinguish different types of play vary based on the number of toys present, the number of people present, as well as the level of play sophistication being recognized. For example, sound and proximity to toys becomes key when trying to determine higher level functional play from imaginary play. If only a single child is present and playing with the toys while a conversation is occurring, higher level functional or imaginary play is likely taking place. For example, a child could instruct the plush puppy rattle to take a running jump over other toys or instruct the plush puppy rattle to have a cup of tea with him. However, if a child and adult are present, conversation may not indicate higher level play. The child may be conversing with the adult rather than speaking to the plush puppy rattle toy or an imaginary play partner. Proximity to the toy may or may not be useful in these cases as the child may clutch the toy while verbally interacting with the adult. Proximity, however, is very useful when distinguishing early exploratory actions from accidental bumps and kicks (see Section 8.4 and Appendix F for more details). In fact, at very early ages, almost all purposeful interactions with objects are considered playful. Using proximity to filter purposeful interactions from accidental interactions is therefore very beneficial when determining early exploratory play from non–play as duration of motion often does not provide enough information to distinguish between the two.

The recognition of object play is further complicated by the fact that higher level object play often mimics real–world activities in pretend play, such as cooking, feeding a baby, cleaning, and shopping. Recognition of general activities of daily living is an open challenge and, in the case of object play, is further complicated because the activities to recognize are performed by young children. The playful nature of children introduce variations in the consistency of how activities are preformed as well as variations due to motor skills development. As mentioned in Section 6.6 the variation inherent in how children preform the same activities on a daily basis and as they develop can make recognition of children's data

more difficult than recognizing similar activities within adult data. Section 8.5 discusses how unsupervised learning algorithms can be used to address this difficulty in future systems.

8.2 Large Scale Data Collection

Long term, wide spread data collection will provide a more representative sample of object play data and may help in the generation of more robust models of object play. In addition to helping researchers analyze play data collected in laboratory studies, the *Child'sPlay* system has been designed to collect and analyze play data within home environments and integrate with systems designed to help parents monitor developmental progress [38, 37]. However, due to social economic factors, it is not realistic to assume that all families can afford such technology. To aid in the collection of large scale data sets as well as help *Child'sPlay* be accessible by all new parents, the *Child'sPlay* system could be integrated into a kiosk and deployed in the waiting areas of pediatricians' offices. Under this scheme, toddlers can play with a set of augmented toys and have their activities characterized while waiting to see the pediatrician during well child visits. In addition to the kiosk being able to upload anonymous play data to a central repository, parents could also keep the results and share them with the pediatrician if there are any concerns. Furthermore, if this centralized model is successful, a similar mechanism can be used to deploy the *Child'sPlay* system in developing countries. *Child'sPlay* characterizes early play motions which psychologists believe to be uniform across many cultures. Therefore, this system has the potential for worldwide deployment and may help assist in the early identification of children with developmental delays in areas where autism awareness is low. In addition to helping identify children and ensuring they receive needed services, deployments of this nature can also help psychologists collect, large scale, multicultural data and further assist with early detection and identification research.

8.3 Selecting Toys for Recording Object Play

Chapter 4 discusses the design and implementation of seven augmented toys designed to record object play data. These toys were designed to interact with each other to help promote early exploratory, relational, and functional object play activities. When used in

clinical settings, it is feasible to use all of the toys. However, seven augmented toys may be impractical for home deployments, due to issues of battery maintenance and toy expense.

Subsets of augmented toys can easily be selected for home deployments. If only one toy can be deployed, a toy with social properties, such as the smiling face on the plush puppy rattle, is recommended. The plush puppy rattle can be used in lower level play and supports a wide variety of higher level play, such as having the puppy run and jump over objects as well as imaginary interactions with the puppy. When collecting the various toddler data sets, the plush puppy rattle toy appeared more approachable by the toddlers due to the social face and familiar dog shape. In addition, the social aspects and familiar shape of the plush puppy rattle toy help promote the detection of appropriate and inappropriate play. For example, if the toy is oriented with the legs pointed towards the ground and moves in the direction that the head is facing, the puppy is likely being used in an appropriate manner. However, if the puppy is dropped face first, repeatedly, there is potential that the child does not recognize the significance of the face or the plush puppy rattle's resemblance to a living creature. The more abstract toys, such as the plastic dome toys, cannot be used in this manner to detect inappropriate play.

Another practical augmented toy subset is a three toy combination consisting of the plush puppy rattle and the two plastic dome toys. The plastic dome toys were specifically designed to support relational play by promoting stacking and assembling of the plastic domes. In addition to forming a ball, the two plastic domes were also designed to allow nesting the plush puppy rattle inside to support a variety of developmentally relevant functional and imaginary play activities. Either one of the domes can serve as an imaginary vehicle for the puppy, such as a car driving across the ground or a spaceship flying through the air. The puppy can also be nested inside the ball and rolled across the floor — this action was seen frequently in both adult freeplay sessions as well as toddler and child play sessions. This three toy combination nicely promotes both the low level exploratory object play as well as the higher level functional and imaginary object play.

If the Lego^{TM} Quatro toys are added to the subset, creating a subset of five toys, the additional toys provide more opportunities for relational play to be recorded. The two

Lego^{TM} Quatro toys also provide another object, in addition to the plush puppy rattle, that can be nested inside the domes. The two Lego^{TM} Quatro toys also provide additional objects for the plush puppy rattle to push around or interact with during higher level object play. The plastic ring would be the next toy to add to the subset, promoting additional higher level play between the puppy and the ring as well as opportunities for relational play between the plastic ring and the plastic dome toys.

When selecting subsets of augmented toys, it is important to select toy combinations that maximize the occurrence of the play activities a researcher or parent wishes to observe. The subset combinations presented above are selected to maximize the observation of differing levels of play sophistication and play type.

8.4 Development of "Smarter" Toys

During most of the data collection described in this dissertation, the augmented toys were used in relatively isolated settings, typically involving a single child playing with them. When siblings played with the toys, or when parent and child interacted with the toys, it was difficult to automatically attribute which motions to associate with which participants. Being able to identify parent from child, or siblings from each other, may help support research involving social aspects of object play. Toy form factor, capacitive sensing, and force resistive sensing offer a potential solution to this issue. Preliminary work suggests that the form factor of a toy, specifically its physical affordances, can cause adults to grasp and manipulate toys in a way that is distinct from young children. The selective placement of capacitive and force resistive sensors in areas where an adult is most likely to grasp the object can provide valuable information to pattern recognition systems [22].

In addition to distinguishing between play participants, capacitive sensing can be important in filtering play motions from accidental bumps and kicks that can occur during play. The *Child'sPlay* system assumes that a child is focusing his attention on the toy that is actively experiencing motion. However, playtime can be very chaotic where toys are unintentionally toppled, kicked, and knocked down. Furthermore, the sensors within the augmented toys picks up vibrations from a child moving around the play space even

when the toys are at rest. Determining the proximity of the child to the toys may help distinguish between intentional and unintentional interactions. The plush cube toy was originally designed to support capacitive sensing for this purpose. However, the capacitive fabric design proved flawed and did not withstand sixty play sessions. Appendix F includes details on the plush cube design as well as new modifications to the Lego^{TM} Quatro design to support robust capacitive sensing and the associated preliminary data.

Capacitive furs and polymers offer the potential to study object play in new ways. For example, it can be difficult for an observer to determine the instant when a child comes in contact with a toy or the moment he releases a toy from video footage. Understanding the way a child reacts to the textures of a toy when he touches it can be instrumental in diagnosing if the child has tactile sensitivities to specific textures [8]. Being able to accurately detect the onset of a grasp or the release of a toy in a quantitative way can help in determining such aversions. Toys augmented with capacitive polymers can help support this type of quantitative analysis and may allow developmental psychologists a new way to explore object play and sensory aversion. In general, augmented toys offer the potential for researchers to sense and visualize aspects of play the human eye cannot and may allow a microscopic analysis of object play that was not previously possible.

The combination of sensors and toy form factors is vast. Toys can be tailored to detect specific play behaviors, or they could remain more general depending on the goals of the researchers. The sensing abilities can also be altered depending if they will interact with other augmented toys or off–the–shelf toys. In home deployments where a smaller subset of augmented toys might be used, additional sensors may be added to help determine when the augmented toys interact with off–the–shelf toys. For example, in the situation where an off–the–shelf toy is being nested inside the plastic dome toy, from a motion standpoint, it can appear as if the augmented plastic dome toy was bumped as the system has no knowledge of the motion imposed upon the plastic dome by the off–the–shelf toy. Light sensors and microphones may help in distinguishing when regular toys interact with the augmented toys versus when the augmented toys are accidentally bumped. An internal array of microphones could help localize points of contact when an augmented toy is bumped

or when it interacts with other off–the–shelf toys. Likewise, light sensors may be able to provide additional information that can help determine if an off–the–shelf toy has been nested inside an augmented toy or if the augmented toy is interacting with the off–the–shelf in some way.

8.5 Adapting Algorithms

Chapter 5 and Chapter 6 show a comparison of recognition results over a combination of normal and augmented toys compared to just augmented toys. While using augmented–only toys helps increase recognition rates, it is desirable to have a system that can accurately recognize object play using a mixture of both types of toys. Improving the sensing capabilities of the augmented toys may help increase recognition rates. However, I feel the combination of computer vision techniques, such as those explored by Wang *et al.*, audio analysis, and augmented toys can vastly increase the recognition of object play as well as expand the types of play that can be automatically recognized. Computer vision can help provide general categorical information such as social play, and the augmented toys may help provide nuanced information about that play. The combination of computer vision, audio analysis, and augmented toys may be the key in recognizing higher level symbolic play. Audio and vision analysis can help determine when a child is speaking to an imaginary play partner or a parent present in the scene.

In addition to detecting a wider variety of play, unsupervised techniques may also be used to help determine variations within play activities as a child develops. For example, as a child's motor coordination develops, the ways in which he manipulates objects will change. While a supervised method can be used to classify play activities into a single group, such as shaking, unsupervised clustering methods can be used to categorize the variations within the group. Variations within shaking may cluster into even and uneven shaking or may be further refined. When investigating data from a single child, the presence of fewer clusters for low level play may indicate more advanced development as there is less variation in how the activity is preformed. Likewise, as the child is performing more sophisticated play, an increase in the number of clusters may indicate development of new

skills. When clustering data between children, larger numbers of clusters may indicate the variety in which play can occur for certain activities. These clustering techniques may allow developmental researchers new ways to investigate phenomena that are difficult, if not impossible for them to observe from video alone. Along those same lines, discovery algorithms may help the system adapt as a child learns new skills for systems that are deployed in home environments.

8.6 Studying User Behavior During Retrospective Review in More Detail

Chapter 7 reported on novice use of the *PlayView* intelligent interface to support retrospectively identifying three specific object play behaviors. Although the novices were trained to use an optimal search strategy, several of them chose to use different strategies. The utility of the *PlayView* interface and other tools designed to support retrospective review of play behaviors is likely dependent on the search strategy employed by the participant and how useful he considers the supplied annotations.

The *PlayView* interface is designed to be flexible to support both expert and novice use. It would be very interesting to observe the types of strategies used by experts when using the *PlayView* interface and the value they placed in computer supplied annotations. In speaking with a developmental psychologist [4], she indicated that she has often left current annotations tools in fast forward and stopped video playback when she sees a behavior of interest occur. Such a search strategy might be considered annotation independent, and it would be interesting to determine if the quality of annotation has an impact with such a strategy. In our studies four novice participants reported using the video heavily to identify when play was occurring and using the computer provided labels as a way to know when to slow down the fast forwarding of the video. In strategies similar to this one, the type of annotation may not matter, just the fact that it is present.

In addition to search strategies, there are different types of annotations strategies. There are exhaustive, interval, and segment based coding strategies. In exhaustive annotation the video is ascribed a label or code every time activity in the video changes. Interval coding ascribes a label for a fixed duration of time based on the predominant activity during that

interval. For example if the duration of the interval was 25 seconds, a single label would be applied for the entire 25 seconds. Segment based coding is similar to spot checking. A video is played for a specified number of seconds and paused, if at the end of that duration, the behavior of interest is displayed, that particular moment in time is labeled as an event.

The *Child'sPlay* system and *PlayView* interface were designed to support exhaustive annotation. It would be interesting to compare performance results based on different types of coding and search strategies using experts.

CHAPTER IX

CONCLUSION

This dissertation has provided evidence to support the thesis that ubiquitous sensing technology and statistical models can be used to help researchers identify object play behaviors collected in naturalistic environments. Furthermore, despite inaccuracies in recognition, the technology described in this dissertation can help reduce the perceived effort of annotating object play data and increase the percentage of play examples that a researcher can view.

In Chapter 4, Chapter 5, and Chapter 6, I showed that wireless sensors embedded in toys can be used to collect data that when modeled can provide automatic characterizations of certain exploratory, relational, and functional play behaviors in both children and adults. As discussed in Chapter 6, I developed a play procedure for use with augmented toys to collect object play data. I then showed that statistical techniques, such as support vector machines, can be used to model object play data. On average, these models can obtain an effective retrieval score of 58.8% with play events that occur over a range of play sophistication for a single child using models constructed from adult play data. These models, while not perfect, still promote an increase in performance when used with the *PlayView* interface to support retrospective review of play data. In Chapter 7, I showed that models which matched these current recognition capabilities allow users to record an increased percentage of play activities when compared to standard practices and that effective retrieval rates also increase. Furthermore, the percentage of play activities logged did not vary significantly when comparing current and future recognition capabilities. However, users showed preference towards higher quality annotations when compared to annotations provided by current recognition capabilities. Based on these findings, I affirm my hypothesis that sensors embedded in objects can provide sufficient data for automatic recognition of certain exploratory, relational, and functional object play behaviors in semi-naturalistic environments and that a continuum of recognition accuracy exists which allows automatic

indexing to be useful for retrospective review.

This dissertation has provided the initial foundation for research augmenting toys and applying statistical models to automatically characterize object play. Moving forward, there are three important areas to explore. The first area involves improvements to the augmented toys and long term, *in situ* data collection. The second area involves algorithmic and external sensing modality enhancements for detecting a wider variety of object play. The third area involves further investigation into expert use of the *PlayView* interface to support retrospective review of play activities. With the prevalence of developmental delay in the United States for young children at approximately ten percent, developing technology to help track developmental progress automatically and promote the early identification of these children is paramount.

APPENDIX A

TERMS AND DEFINITIONS

1. **Augment:** The concealment of wireless sensors within a toy. These sensors are capable of detecting acceleration, touch, sound, and limited proximity.

2. **Toys:** Age appropriate objects that can easily be grasped and manipulated by the child. These objects must be large enough to permit augmentation.

3. **Detection:** The automatic identification of child-object interactions from multiple streams of time-series data generated by manipulation of augmented toys.

4. **Interactions:** Any purposeful contact, incidental contact or manipulation of augmented toys.

5. **Categorize:** Interactions will be grouped according to an object-play scale developed by Baranek *et al.* [7]. This scale has quantized exploratory, relational, and function play into twelve levels. Due to limitations imposed by sensor technology, this work will group interactions according to levels zero through six. Higher levels of Baranek's scale is beyond the scope of this work.

6. **Exploratory Play:** Any child's action upon a *single object* that results from a visually–guided reach and helps provide information about the object or environment. No functional relations exist between action and objects. Examples include: (Level 1) grasping, rubbing, shaking, scratching, banging, poking, mouthing, (Level 2) rolling a car, pushing a button, rocking a horse, and opening/closing doors.

7. **Relational Play:** When *two or more objects* are used in combination with each other but are associated without regard to the functions or attributes of the objects. Examples include: (Level 3) pushing apart pop-beads, removing lids from containers, (Level 4) stacking blocks, detaching puzzles pieces, and scooping/pouring objects.

8. **Functional Play:** Any conventional use of an object influenced by cultural properties of the object and simple pretend play actions. Examples include: (Level 5) placing a lid on a pot, dumping objects from a truck, (Level 6) drinking from an empty cup, and raising a phone to an ear to talk to a pretend friend.

9. **Symbolic Play:** Any scheme in a continuum of play schemes that incorporates items, attributes, contexts not actually present, or the substitution of objects. Examples include: (Level 9) using a block as a car, or banana as a phone, (Level 10) using figures to load objects into truck, propping a bottle in a doll's arms to feed her, (Level 11) pretending a doll is crying, or claiming a toy stove is hot to the touch.

APPENDIX B

EVALUATION METRICS FOR CONTINUOUS RECOGNITION

B.1 Levels of Analysis for Continuous Recognition

A fundamental issue for evaluating activity recognition concerns the level of analysis used to calculate performance. Figure 43 illustrates three levels of analysis: event–based, frame–based, and segment–based correspondence.

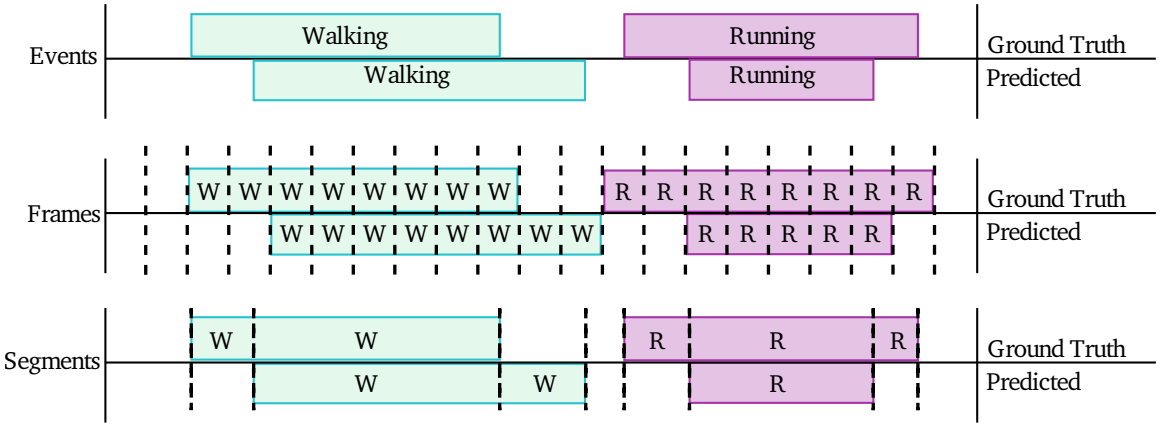


Figure 43: Three levels of analysis: events, frames, and segments.

B.1.1 Event Level Analysis

Each occurrence of an activity represents one *event*, which is a contiguous block of time during which the activity label is constant. Evaluation could measure whether each event is detected, whether the event is detected at the correct time, and how closely the predicted event boundaries correspond to the true start and stop times. We call evaluation at this level *event analysis*. See the top row of Figure 43 for an illustration of events detected at the correct time but with poor boundary alignment (see Appendix B.2 for the details of temporal correspondence of events).

B.1.2 Frame Level Analysis

Alternatively, we can view the data as a series of equal-length time intervals and consider the activity being performed during each interval. For example, we could divide an hour of data into 3,600 seconds and label each one-second block according to the dominant activity during that time. Evaluation would then depend on the correspondence between the ground truth label and predicted label for each second. We call evaluation at this level *frame analysis*. Note that the temporal duration of a frame is arbitrary and can change based on the domain. See the middle row of Figure 43 for an illustration of frame analysis.

B.1.3 Segment Level Analysis

Finally, a hybrid approach divides the data into variable length segments. The segments are defined as maximal intervals within which both the predicted and true labels are constant. Thus the boundary of each segment coincides with a boundary of either a true or predicted label. Evaluation at this level is called *segment analysis* [71]. Thus, each segment may have a different duration, but there are no aliasing problems or ambiguities associated with event correspondences and boundary alignment (see Figure 43). Segment-based representations simplifies detecting different kinds of recognition errors.

B.2 Temporal Correspondence and Identification at the Events Level of Analysis

Each occurrence of an activity represents one *event*, which is a contiguous block of time during which the activity label is constant. Evaluation could measure whether each event is detected, whether the event is detected at the correct time, and how closely the predicted event boundaries correspond to the true start and stop times. We call evaluation at this level *event analysis*. See the top row of Figure 43 for an illustration of events detected at the correct time but with poor boundary alignment.

When it is important to consider the actual time during which an event was detected, a temporal correspondence method can be used. This method seeks to match each ground truth event with a predicted event based on temporal overlap. Many different match criteria are possible (see Figure 44), including:

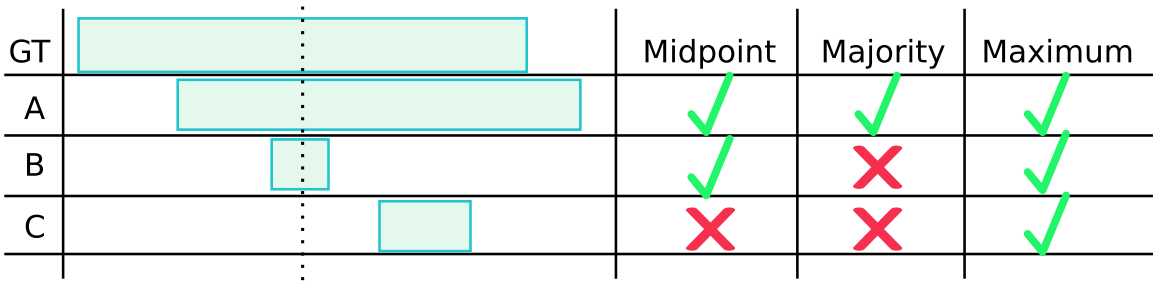


Figure 44: Three methods for determining temporal event correspondence: midpoint span, majority vote, and maximum overlap. The vertical, dashed line represents the midpoint of the ground truth label.

midpoint overlap: A predicted event must span the midpoint of its matching ground truth event. This approach is often used to score word spotting systems in the speech recognition domain.

majority overlap: A predicted event is paired with a ground truth event if the overlap accounts for a majority of the time in both events.

maximum overlap: Predicted and ground truth events are paired based on maximizing overlap. Although computing the optimal correspondence is NP-hard, greedy approaches work well in practice.

B.3 Types of Errors Encountered in Continuous Recognition

There are many types of errors that can occur during continuous recognition involving correspondence issues between activity boundaries and labels. Figure 3 shows the output of nine different recognition systems, $A-J$ where each illustrates a specific error type common to continuous recognition [50]. These error types are:

Correct (C): sometimes called "Hits" (H); represents correct classification. This number represents both *True Positives* and *True Negatives* (see Figure 3, System A)

Substitutions (S): represent correct temporal detection but incorrect activity identification (see Figure 3, System B)

Insertions (I): detection of an activity when none actually occurred; this can also occur when a long activity is partially detected multiple times (see Figure 3, System C)

Deletions (D): failure to detect an activity (see Figure 3, System D)

Total Number of True Events (N): a useful variable for calculating statistics, though not strictly a classification result: $N = (C + D + S)$.

Underfill (U): when an activity is correctly identified, underfill errors account for the time at the beginning and end of the activity that is not detected (see Figure 3, System E)

Overfill (O): when an activity is correctly detected, overfill errors account for the time before and after the activity that is incorrectly identified as part of the activity (see Figure 3, System F)

Fragmentation (F): errors due to detecting a long activity as multiple events separated by null(see Figure 3, System G)

Substitution-fragmentation (S_F): whereas a normal fragmentation error falsely divides an event with null, this error divides an event by incorrectly inserting known activities (see Figure 3, System H)

Merge (M): errors due to incorrectly detecting multiple, closely occurring events that are separated by null as a single, longer event (see Figure 3, System I)

Substitution-merge (S_M): like a standard merge error, a substitution-merge involved detecting multiple occurrences of an activity as a single occurrence, but here the separating activity is a known class (see Figure 3, System J)

B.4 Common Evaluation Metrics

Once the different continuous recognition error types have been accumulated, a variety of summary statistics can be computed. Each statistic highlights the recognition system's performance relative to different criteria. Many of the most commonly used metrics are presented below [51, 50].

B.4.0.1 Sensitivity / Recall

Sensitivity, which is also referred to as recall, corresponds to the correct detection rate relative to ground truth. It is the percentage of correctly detected activities out of all true instances of a particular class, averaged over all activities. Sensitivity is defined as $\frac{TP}{TP+FN}$. Likewise, $(1 - sensitivity) = \frac{FN}{TP+FN}$ is the probability of the recognizer failing to detect an instance of an activity.

B.4.0.2 Precision / Positive Predictive Value

Precision is also known as the Positive Prediction Value (PPV) and measures the likelihood that a detected instance of an activity corresponds to a real occurrence. Precision is defined as $\frac{TP}{TP+FP}$. Likewise, $(1-precision) = \frac{FP}{TP+FP}$ is the probability of the recognizer incorrectly identifying a detected activity.

Precision and recall are highly related. Both are based on the number of true positives, but sensitivity normalizes by the true number of occurrences (based on the ground truth), while recall normalizes by the total number of occurrences detected (based on the predicted label). Thus, they estimate different likelihoods: "What percentage of the total number of occurrences will the recognizer correctly identify?" and "What percentage of the detected occurrences will be correct?".

B.4.0.3 Specificity

Specificity can be thought of as the recognizer's sensitivity to the negative class. It measures the proportion of correctly identified negative occurrences to all true negative occurrences. Specificity is defined as $\frac{TN}{TN+FP}$

B.4.0.4 Negative Predictive Value (NPV)

The negative predictive value can be thought of as "negative precision" and measures the likelihood that a negative identification is correct relative to all negative identifications. The negative predictive value is defined as $NPV = \frac{TN}{TN+FN}$.

B.4.0.5 F-Measure

The F-Measure combines the precision and recall rates into a single measure of performance. It is defined as the harmonic mean of precision, P, and recall, R: $F = \frac{(\beta^2+1)PR}{\beta^2 P+R}$, where precision and recall are evenly weighted when $\beta = 1$ [58, 30].

B.4.0.6 Likelihood Ratio

The Likelihood Ratio is the ratio of the likelihood that a particular activity would be predicted when it matches the ground truth to the likelihood that it would be predicted erroneously. This ratio can be computed for both true positive and true negative results:

$$LR+ = \frac{sensitivity}{1-specificity} = \frac{TP(TN+FP)}{FP(TP+FN)}$$

$$LR- = \frac{1-sensitivity}{specificity} = \frac{FN(TN+FP)}{TN(TP+FN)}$$

B.4.0.7 Accuracy

Accuracy is defined as $\frac{(C-I)}{N}$ and measures the percentage of correct identifications after discounting insertion errors. Although accuracy has a maximum value of 100%, for continuous recognition systems, there is no general lower bound due to the penalty for insertion errors. Thus, a poor recognition system with a very low detection threshold could insert more false detections than there are true events, thereby leading to a negative overall accuracy.

APPENDIX C

GT^{2}K MATHEMATICAL DETAILS

C.1 HMM Parameter Learning used in GT^{2}k

Training of an HMM involves adjusting the model parameters $\lambda = (A, B, \lambda)$ to maximize the probability of generating the observation sequence (in our case specific sensor readings) given the model λ. There is no way to analytically solve for λ which maximizes $P(O|\lambda)$ [57]. Parameters for λ can be computed such that they locally optimize $P(O|\lambda)$ by an iterative method known as Baum–Welch re–estimation. In this process the initial model $\lambda = (A, B, \pi)$ is used in the re–estimation equations (see equations 6, 7, and 8 given below) to produce a new estimate $\overline{\lambda} = (\overline{A}, \overline{B}, \overline{\pi})$. We then iteratively use $\overline{\lambda}$ in place of λ which allows us to improve the probability of O being observed from the model. After each iteration two possible conditions hold: the model λ defines a critical point and $\overline{\lambda} = \lambda$, or the model $\overline{\lambda} = \lambda$ is more likely to have produced the observation sequence than the model λ such that $P(O|\overline{\lambda}) > P(O|\lambda)$. Re–estimation is repeated until λ defines a critical point or until some limiting point is reached.

The model parameters can be re–estimated using frequency counting [57]. Equation 8 is the re–estimation of the observation probability density and Equation 7 is the re–estimation of transition probability.

$$\overline{\pi}_i = \gamma_1(i) = expected\ number\ of\ times\ in\ S_i\ at\ t = 1 \tag{6}$$

$$\overline{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} = \frac{expected\ number\ of\ transitions\ from\ S_i\ to\ S_j}{expected\ number\ of\ transitions\ from\ S_i} \tag{7}$$

$$\overline{b}_j(k) = \frac{\sum_{t=1\ O_t=v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T-1} \gamma_t(i)} = \frac{expected\ visits\ to\ state\ S_i\ and\ observing\ symbol\ v_k}{expected\ number\ of\ times\ in\ S_j} \tag{8}$$

C.2 Recognition used in GT^{2}k

We computed the probability of an observed sequence, $O =< o_1, o_2, o_3, \ldots, o_T >$ being generated by a specific model, $P(O|\lambda)$, using the Viterbi Algorithm. Each observation O is associated with a state S_J given a model λ. Thus, given an observation sequence O and a model λ, $P(Q, O|\lambda)$ is computed by determining the a single path through the model (the state sequence $Q =< q_1, q_2, \ldots, q_T >$) that best explains the observation sequence. This quantity is calculated through dynamic programing methods [1].

In isolated recognition, the observation sequence O is aligned to at most one model. In continuous recognition, the observation sequence O may align with multiple models (because it contains multiple activities). Alignment is performed by constructing the most probable path through all possible sequences of the models in parallel. Domain knowledge, in the form of simple grammars, can be used to help inform the alignment process[2]. Grammars, by providing *a priori* activity sequence knowledge, help prune the search space by eliminating model sequences that violate the grammar. This reduces both alignment time and misalignment with improbable sequences.

[1] Mathematical formulation and implementation details can be found in Rabiner's tutorial [57]

[2] There has been much literature in the speech–recognition domain and the mathematical details of this process can be found in [29]

APPENDIX D

PILOT ALGORITHM MATHEMATICAL DETAILS

D.1 Aggregation of Features for Simple Spacial Recognition

Several steps are needed to prepare the raw accelerometer readings for analysis. First, we resample each sensor stream at 60Hz to estimate the instantaneous reading of every sensor at the same fixed intervals. Next, we slide a three second window along the 18D synchronized time series in one second steps. For each window, we compute 378 features based on the 18 sensor readings including mean, variance, RMS, energy in various frequency bands, and differential descriptors for each dimension. We also compute aggregate features based on each three-axes accelerometer including the mean, variance, and RMS of the magnitude of the sensor reading in 3-space and based on the angle of the vector relative to the x-axis.

The computation of aggregate features transforms a difficult temporal pattern recognition problem into a simpler spatial classification. Rather than explicitly choosing relevant features from the aggregate set as a preprocessing step, models are built using an adaboost framework [60] by selecting the best dimension and 1D classifier during each iteration. This framework automatically selects the best features for discrimination and, importantly, is robust to unimportant or otherwise distracting features.

D.2 Boosting One–Dimensional Classifiers

Adaboost is an iterative framework for combining binary classifiers such that a more accurate ensemble classifier results [60]. We use a variant of the original formulation that includes feature selection and support for unequal class sizes [67]. Given a training data set $(x_1, y_1), ..., (x_n, y_n)$, where each pair (x_i, y_i) consists of a feature vector ($x_i \in \mathcal{R}^N$) and a label ($y_i = \pm 1$), each round of boosting selects the dimension and 1D classifier that minimizes the classification error over the weighted training set. Initially, the weights are set to be uniform within each class: $w_i^1 = [\frac{1}{2p}\ \textit{if}\ y_i = +1\ \textit{else}\ \frac{1}{2q}]$, where p is the number of

positive examples and q is the number of negative examples. In each round of boosting (m), the weights are normalized ($w_i^m \leftarrow \frac{w_i^m}{\sum_{j=1}^n w_j^m}$), and a new weak classifier is selected by searching over all weak learners ($h_j(x)$) and all features and then choosing the classifier ($h^m(x)$) which minimizes the weighted error ($\epsilon_j^m = \sum_i w_i^m I(h_j^m(x_i) \neq y_i)$), where $I(\cdot)$ is the indicator function that equals one when the condition is true and zero otherwise.

After the best feature and 1D classifier is found, the weights are updated according to $w_i^{m+1} = w_i^m \beta_m^{I(h^m(x_i) \neq y_i)}$, where $\beta_m = \frac{\epsilon^m}{1-\epsilon^m}$. Boosting continues for a specified number of iterations (M), and then the final ensemble classifier ($H(x)$) is formed as a weighted combination of the weak classifiers: $H(x) = sign(\frac{\sum_{m=1}^M \alpha_m h^m(x)}{\sum_{m=1}^M \alpha_m})$, where $\alpha_m = log(\frac{1}{\beta_m})$.

The ensemble classifier is based on the sign of a value derived from a weighted combination of the weak classifiers called the margin ($m(x) = (\frac{\sum_{m=1}^M \alpha_m h^m(x)}{\sum_{m=1}^M \alpha_m})$. The magnitude of the margin gives an indication of the confidence of the classifier in the result. Margins may not be comparable across different classifiers however, but we can use a method developed by Platt to convert the margin into a probability [55]. This method works by learning the parameters to a sigmoid function that directly maps from a margin to the probability of one of the classes (without loss of generality, we take this as $p(y = +1) = p(\omega_1)$). Since $p(\omega_1|x) = \frac{p(x|\omega_1)p(\omega_1)}{\sum_j p(x|\omega_j)p(\omega_j)}$, it suffices to estimate $p(x|\omega_j)$, which we can do via kernel density estimation after computing the margin for each training point. Finally, we fit a sigmoid ($f(x) = 1/(1 + e^{A(x+B)})$) to the margin/probability pairs derived from the training points and only save the sigmoid parameters (A and B) for use during inference.

D.3 Selection of One–Dimensional Weak Classifiers

During each round of boosting, the algorithm selects the feature and 1D classifier that minimizes the weighted training error. Typically, decision stumps are used as the 1D classifier due to their simplicity and efficient, globally optimal learning algorithm. Decision stumps divide the feature range into two regions, one labeled as the positive class and the other as the negative class. In our experiments, we supplement decision stumps with a Gaussian classifier that models each class with a Gaussian distribution and allows for one, two, or three decision regions depending on the model parameters.

Learning the Gaussian classifier is straightforward. The parameters for the two Gaussians ((μ_1, σ_1) and (μ_2, σ_2)) can be estimated directly from the weighted data. The decision boundaries are then found by equating the two Gaussian formulas and solving the resulting quadratic equation for x: $(\sigma_1^2 - \sigma_2^2)x^2 - 2(\sigma_1^2\mu_2 - \sigma_2^2\mu_1)x + [(\sigma_1^2\mu_2^2 - \sigma_2^2\mu_1^2) - 2\sigma_1^2\sigma_2^2 log(\frac{\sigma_1}{\sigma_2})] = 0$.

In general, arbitrarily complex weak learners can be used within a boosting framework. In our case, the resulting ensemble learner always has the same form (axis-aligned decision boundaries) but may learn an accurate model more quickly (*i.e.,* fewer rounds of boosting) or have better generalization depending on the choice of weak learners. We empirically determined that boosting over both decision stumps and Gaussian classifiers led to a sufficient increase in classification accuracy to justify the extra computation during learning. Specifically, when testing on Aberdeen data that included the null activity, cross-validated, event-based accuracy rose from 71.5% to 74.9% when we boosted over both 1D classifiers.

D.4 Combining Binary–class Classifiers for Multi–class Classification

The boosting framework can be used to learn accurate binary classifiers. Two common methods for combining multiple binary classifiers into a single multiclass classifier are the one–vs–all (OVA) and the one–vs–one (OVO) approaches. In the one–vs–all case, for C classes, C different binary classifiers are learned. For each classifier, one of the classes is taken as the positive class while the $(C-1)$ others are combined to form the negative class. When a new feature vector must be classified, each of the C classifiers is applied and the one with the largest margin (corresponding to the most confident positive classifier) is taken as the final classification.

In the one-vs-one approach, $C(C-1)/2$ classifiers are learned, one for each pair of (distinct) classes. To classify a new feature vector, the margin is calculated for each class pair (m_{ij} is the margin for the classifier trained for ω_i vs. ω_j). Within this framework many methods may be used to combine the individual classification results:

vote[ms,ps,pp]: each classifier votes for the class with the largest margin; ties are broken by using *msum*, *psum*, or *pprod*.

msum: each class is scored as the sum of the individual class margins; the

final classification is based on the highest score:

$\underset{i}{argmax} \sum_{j=1}^{C} m_{ij}(x)$

psum: each class is scored as the sum of the individual class probabilities; the final classification is based on the highest score:

$\underset{i}{argmax} \sum_{j=1}^{C} p(\omega_i|m_{ij}(x))$

pprod: the classification is based on the most likely class:

$\underset{i}{argmax} \prod_{j=1}^{C} p(\omega_i|m_{ij}(x))$

D.5 Combining Classification Results from Overlapping Windows

As discussed in Section D.1, the temporal recognition problem is transformed into a spatial classification task by computing features over three second windows spaced at one second intervals. This means that typically there are three different windows that overlap for each second of data. To produce a single classification for each one second interval, three methods were tested:

vote: each window votes for a single class and the class with the most votes is selected

prob: each window submits a probability for each class, and the most likely class is selected: $\underset{i}{argmax} \prod_{j=1}^{3} p(\omega_i|window_j)$

probsum: each window submits a probability for each class, and the class with the highest probability sum is selected: $\underset{i}{argmax} \sum_{j=1}^{3} p(\omega_i|window_j)$

D.6 Implications

The method described in this section is influenced by many parameters. With respect to the task of identifying soldier activities in the field, recognition performance is relatively unaffected by the choice of multi–class aggregation method, though one–vs–one methods (see Section D.4) lead to a considerable reduction in training time. This distinction can be important for classification systems that have large numbers of classes. The voting scheme

for combining classification results, however, can have a noticeable impact on the frame level accuracy rate. Surprisingly, the number of rounds of boosting did have a slight impact on results, however, the results tended to stabilize after 20 rounds.

APPENDIX E

SURVEY PACKETS

Participant ID:_____ Date:_____ Time:_____

Background Information

Please answer the following questions. If you answer "NO" to a question in bold, you may skip the remaining parts, such as (a) (b) (c), for that question. Feel free to ask questions at any time.

1. **What year were you born?** _____

2. **In what country were you born?** _____

3. **What is your current/highest level of education?**(Circle one)

   Ph.D Masters Bachelors Associates High School Other_____

4. **What is your occupation and/or field of study?**_____

5. **How many children do you have?** (List gender and ages)_____

6. **How many hours a week do you use a computer?** (Circle one)

150 + hours (7)	100 – 149 hours (6)	50 – 99 hours (5)	30 – 49 hours (4)	10 – 29 hours (3)	5 – 9 hours (2)	1 – 4 hours (1)

7. **Have you ever used video editing software before?** (Circle one) YES NO

   a. List the software(s) you have used and when you last used them: _____

   b. How many hours have you spent editing video footage? (Circle one)

150 + hours (7)	100 – 149 hours (6)	50 – 99 hours (5)	30 – 49 hours (4)	10 – 29 hours (3)	5 – 9 hours (2)	1 – 4 hours (1)

8. **Have you ever annotated or transcribed video?** (Circle one) YES NO

   a. Which software or method did you use and when did you last use them? _____

   b. What type of video did you annotate and why? _____

   c. How many hours have you spent annotating video footage? (Circle one)

150 + hours (7)	100 – 149 hours (6)	50 – 99 hours (5)	30 – 49 hours (4)	10 – 29 hours (3)	5 – 9 hours (2)	1 – 4 hours (1)

Participant ID:_____ Date:_____ Time:_____

Background Information

Please answer the following questions. If you answer "NO" to a question in bold, you may skip the remaining parts, such as (a) (b) (c), for that question. Feel free to ask questions at any time.

9. **Which of the following courses have you taken?** (Circle **ALL** that apply)

Pattern Recognition	Machine Learning	Computer Vision	Data Mining	Medical Image Analysis	None of the courses listed

a. In these classes, did you ever implemented pattern recognition algorithms? YES NO

b. In these classes, did you ever labeled data for use with these algorithms? YES NO

10. **Have you ever been involved in clinical research involving young children?** YES NO

11. **Have you ever observed young children playing with toys?** (Circle one) YES NO

a. In what type of setting (*ie*: home, lab) _____

b. For what purpose (*ie*: parenting, babysitting, for a study) _____

c. How many hours have you spent observing children play? (Circle one)

150 + hours (7)	100 – 149 hours (6)	50 – 99 hours (5)	30 – 49 hours (4)	10 – 29 hours (3)	5 – 9 hours (2)	1 – 4 hours (1)

12. **Are you color blind or have difficulty distinguishing between different colors?** YES NO

13. **What is your opinion of the following statement?** (Circle one)

"If a computer uses an algorithm to provide me with information, I believe it to be correct."

Strongly Agree (7)	Agree (6)	Somewhat Agree (5)	Neutral (4)	Somewhat Disagree (3)	Disagree (2)	Strongly Disagree (1)

Optional: Are there any additional comments you would like to share with the researcher? _____

Participant ID:_____ Participant Condition:_____ Condition Code:_____

Post Condition Survey

Rankings

Rank the following items in order according to category based on the task you just completed. Numbers may only be used once per category. Comments are optional.

1. **Ease of Identification** (3 = easiest to identify, 1 = hardest to identify)

   ☐ Assembling LEGOS™ ☐ Jumping Puppy over toys ☐ Shaking a toy

   comments:

2. **Time Spent** (5 = majority of my time, 1 = least of my time)

   ☐ Watching video ☐ Creating labels ☐ Adjusting labels ☐ Searching for/over labels ☐ Logging labels

   comments:

3. **Effort**(5 = most difficult, 1 = least difficult)

   ☐ Comprehending play seen in video ☐ Ignoring labels not related to my task ☐ Looking for so many of my play activities ☐ Completing Task on time ☐ Correcting labels

   comments:

Questions

Circle the answer that most accurately reflects your opinion.

4. **I am satisfied with the number of labels provided by the computer**.

Strongly Agree (7)	Agree (6)	Somewhat Agree (5)	Neutral (4)	Somewhat Disagree (3)	Disagree (2)	Strongly Disagree (1)

   a. **If not, where there too many or too few**?

5. **The locations of my play activities were easy to find**.

Strongly Agree (7)	Agree (6)	Somewhat Agree (5)	Neutral (4)	Somewhat Disagree (3)	Disagree (2)	Strongly Disagree (1)

Participant ID:_____ Participant Condition:_____ Condition Code:_____

Post Condition Survey

6. **The beginning and ending of my play activities were not accurately labeled**.

Strongly Agree (7)	Agree (6)	Somewhat Agree (5)	Neutral (4)	Somewhat Disagree (3)	Disagree (2)	Strongly Disagree (1)

7. **I was distracted by labels not related to my task**.

Strongly Agree (7)	Agree (6)	Somewhat Agree (5)	Neutral (4)	Somewhat Disagree (3)	Disagree (2)	Strongly Disagree (1)

8. **Each label I viewed accurately identified what was seen in the associated video**.

Strongly Agree (7)	Agree (6)	Somewhat Agree (5)	Neutral (4)	Somewhat Disagree (3)	Disagree (2)	Strongly Disagree (1)

9. **Erroneous labels prevented me from completing my task.**

Strongly Agree (7)	Agree (6)	Somewhat Agree (5)	Neutral (4)	Somewhat Disagree (3)	Disagree (2)	Strongly Disagree (1)

10. **I found computer generated labels that were useful to my search**.

Strongly Agree (7)	Agree (6)	Somewhat Agree (5)	Neutral (4)	Somewhat Disagree (3)	Disagree (2)	Strongly Disagree (1)

11. **I found the overall task difficult**.

Strongly Agree (7)	Agree (6)	Somewhat Agree (5)	Neutral (4)	Somewhat Disagree (3)	Disagree (2)	Strongly Disagree (1)

Participant ID:_____ Participant Condition:_____ Condition Code:_____

Post Condition Survey

12. **I am confident that I logged all instances of my play activates that exist in the video.**

Strongly Agree (7)	Agree (6)	Somewhat Agree (5)	Neutral (4)	Somewhat Disagree (3)	Disagree (2)	Strongly Disagree (1)

13. **Overall, the computer reduced the amount of effort required to annotate this video**.

Strongly Agree (7)	Agree (6)	Somewhat Agree (5)	Neutral (4)	Somewhat Disagree (3)	Disagree (2)	Strongly Disagree (1)

My Strategy

14. **Please describe the strategy you used to complete this task**:

_____
_____
_____
_____
_____
_____
_____
_____

Additional Feedback

Is there anything else you would like to share with the researchers?:

_____
_____
_____
_____
_____
_____
_____
_____

Participant ID:_____

Post Experiment Survey

Ranking of Conditions

Rank your 4 conditions according to the following items. Please refer to the provided screen shots if you need help recalling details about each condition. Numbers may only be used once per category. Comments are optional -- If you find a ranking difficult to assign, please make a note in the comments.

1. Which condition did you **Like the Best?** (4 = Best, 1 = Worst)

   ☐ Condition A ☐ Condition B ☐ Condition C ☐ Condition D

   Comments: _____

2. Which condition **provided labels that were Least Useful?** (4 = Least useful, 1 = Most useful)

   ☐ Condition A ☐ Condition B ☐ Condition C ☐ Condition D

   Comments: _____

3. Which condition was **Easiest to Find "My Play" Activities** (4 = Easiest, 1 = Most Difficult)

   ☐ Condition A ☐ Condition B ☐ Condition C ☐ Condition D

   Comments: _____

4. Which condition took the **Most Effort?** (4 = Most Effort, 1 = Least Effort)

   ☐ Condition A ☐ Condition B ☐ Condition C ☐ Condition D

   Comments: _____

Participant ID:_____

Post Experiment Survey

Ranking of Label Utility

Rank the following items in order according to category based on ALL the conditions. Numbers may only be used once per category. Comments are optional.

5. **Provided Labels were Most Import for Identifying** (5 = Most important, 1 = Least important)

☐ Type of play ☐ Location of play ☐ Duration of play ☐ Start of play ☐ End of play

Comments: _____

_____

_____

6. It is **Most Important** that the labels Accurately Provide (5 = Most important, 1 = Least important)

☐ Type of play ☐ Location of play ☐ Duration of play ☐ Start of play ☐ End of play

Comments: _____

_____

_____

7. Which had the **Most Influence on Task Completion** (5 = Most influence, 1 = Least influence)

☐ The number of activities to find ☐ Pre-computed labels being provided ☐ Labels providing accurate locations ☐ Labels providing accurate names ☐ Allotted Time

Comments: _____

_____

_____

Labels (Annotations)

8. **There were two conditions which seemed to provide the same quality of labels**.

Strongly Agree (7)	Agree (6)	Somewhat Agree (5)	Neutral (4)	Somewhat Disagree (3)	Disagree (2)	Strongly Disagree (1)

a. **If so, which two conditions were indistinguishable**? _____

Participant ID:_____

Post Experiment Survey

9. **If the computer provides a labeled location, I believe it to be correct.**

Strongly Agree (7)	Agree (6)	Somewhat Agree (5)	Neutral (4)	Somewhat Disagree (3)	Disagree (2)	Strongly Disagree (1)

10. **It is easier to ignore extraneous labels than it is to search video for missing labels.**

Strongly Agree (7)	Agree (6)	Somewhat Agree (5)	Neutral (4)	Somewhat Disagree (3)	Disagree (2)	Strongly Disagree (1)

11. **Overall, the computer accurately identified the play activities I was trying to find**.

Strongly Agree (7)	Agree (6)	Somewhat Agree (5)	Neutral (4)	Somewhat Disagree (3)	Disagree (2)	Strongly Disagree (1)

12. **I would rather have inaccurate computer generated labels than no labels at all.**

Strongly Agree (7)	Agree (6)	Somewhat Agree (5)	Neutral (4)	Somewhat Disagree (3)	Disagree (2)	Strongly Disagree (1)

13. **Computer generated labels increased the amount of effort required to annotate video**.

Strongly Agree (7)	Agree (6)	Somewhat Agree (5)	Neutral (4)	Somewhat Disagree (3)	Disagree (2)	Strongly Disagree (1)

14. **I relied solely on annotations provided by the computer to complete my task**.

Strongly Agree (7)	Agree (6)	Somewhat Agree (5)	Neutral (4)	Somewhat Disagree (3)	Disagree (2)	Strongly Disagree (1)

15. **Computer generated labels made it difficult to search for multiple types of play at once.**

Strongly Agree (7)	Agree (6)	Somewhat Agree (5)	Neutral (4)	Somewhat Disagree (3)	Disagree (2)	Strongly Disagree (1)

Participant ID:_____

Post Experiment Survey

The Interface and Video Navigation

16. **The interface was easy to use**.

Strongly Agree (7)	Agree (6)	Somewhat Agree (5)	Neutral (4)	Somewhat Disagree (3)	Disagree (2)	Strongly Disagree (1)

17. **It was difficult to move from one annotation to the next**.

Strongly Agree (7)	Agree (6)	Somewhat Agree (5)	Neutral (4)	Somewhat Disagree (3)	Disagree (2)	Strongly Disagree (1)

18. **Annotations are easy to create once I identify the location in the video I wanted to mark**.

Strongly Agree (7)	Agree (6)	Somewhat Agree (5)	Neutral (4)	Somewhat Disagree (3)	Disagree (2)	Strongly Disagree (1)

19. **The interface hindered me from completing my task**.

Strongly Agree (7)	Agree (6)	Somewhat Agree (5)	Neutral (4)	Somewhat Disagree (3)	Disagree (2)	Strongly Disagree (1)

20. **I made use of more than one camera view**.

Strongly Agree (7)	Agree (6)	Somewhat Agree (5)	Neutral (4)	Somewhat Disagree (3)	Disagree (2)	Strongly Disagree (1)

21. **The quality of the video made completing my task difficult.**

Strongly Agree (7)	Agree (6)	Somewhat Agree (5)	Neutral (4)	Somewhat Disagree (3)	Disagree (2)	Strongly Disagree (1)

Additional Feedback

Is there anything else you would like to share with the researchers?

_____

_____

_____

_____

APPENDIX F

POSTER PAPERS ON CAPACITIVE SMART TOYS

Capacitance Sensing in Smart Toys: Aiding the Detection of Play Behaviors

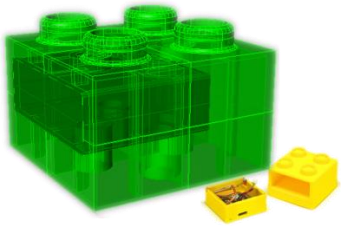


Figure 1: Plastic block CAD model

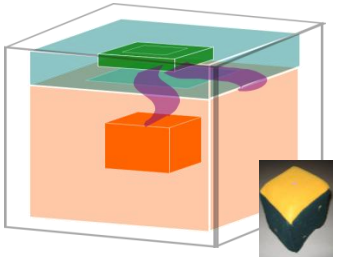


Figure 2: X-ray view of the plush cube. The upper region (green) is the sensor package inside a foam shelf. The lower region (orange) is concentric cubes covered in conductive fabric. Wires (purple) connect the inner and outer conductive shells to the sensor.

Tracy L. Westeyn
College of Computing, GVU
Georgia Institute of Technology
85 5th Street, NW
Atlanta, GA 30332 USA
turtle@cc.gatech.edu

Jiasheng He
College of Computing, GVU
Georgia Institute of Technology
85 5th Street, NW
Atlanta, GA 30332 USA
jiazhe@cc.gatech.edu

Peter W. Presti & Jeremy M. Johnson
Interactive Media Technology
Georgia Institute of Technology
85 5th Street, NW
Atlanta, GA 30332 USA
peter.presti@imtc.gatech.edu
johnsonj@imtc.gatech.edu

David Quigley
College of Computing, GVU
Georgia Institute of Technology
85 5th Street, NW
Atlanta, GA 30332 USA
quigs@gatech.edu

Scott Gilliland
College of Computing, GVU
Georgia Institute of Technology
85 5th Street, NW
Atlanta, GA 30332 USA
scott.gilliland@gatech.edu

Thad E. Starner
College of Computing, GVU
Georgia Institute of Technology
85 5th Street, NW
Atlanta, GA 30332 USA
thad@cc.gatech.edu

Copyright is held by the author/owner(s).
UbiComp 2009, Sep 30 – Oct 3, 2009, Orlando, FL, USA

Abstract
Our recent research has investigated the use of wireless accelerometers embedded in toys to aid in the automatic detection and analysis of children's playtime activities. This paper discusses the implementation and sensing capabilities of two augmented toys, a plush cube and a Lego™ Quatro compatible block. One goal of these toys is to distinguish between a child's direct manipulations as opposed to motions caused by kicks, accidental bumps, and other indirect interactions.

Keywords
Activity Recognition, Toy Design, Object-Play, Multimodal Wireless Sensing

ACM Classification Keywords
H3.1. Content Analysis and Indexing;

Introduction & Motivation
The way in which infants play with objects can serve as an early indicator for developmental delays [1]. We designed several smart toys to aid in the automatic detection and analysis of children's playtime activities [3]. Typically, a toddler focuses his attention on the toys he is actively manipulating; however, the toys in his hands may not be the only toys experiencing motion. Playtime can be a very chaotic activity where toys are unintentionally toppled, kicked, and knocked down. Often, acceleration data does not provide enough information to determine which toys are held in the hands and which toys are accidentally bumped. This

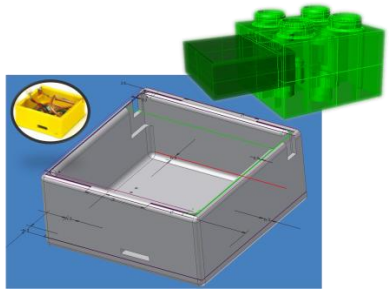


Figure 3: Plastic block and drawer CAD models. Peepholes exist in the back corners of the left and right walls.

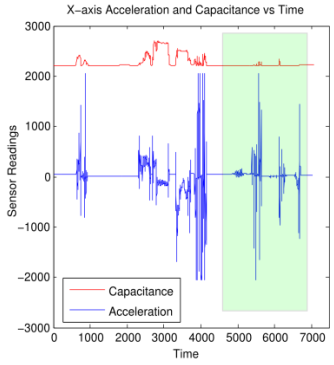


Figure 4: The green region highlights where the plastic block was indirectly manipulated. The upper (red) line is capacitance while the lower line (blue) is acceleration.

distinction is important for automatically generating quantitative measures of object play sophistication. Detecting the proximity of the child through capacitive sensing may help distinguish between the intentional and unintentional interactions.

Toy Designs & Discussion

Our toy designs leverage the *BlueSense* sensor package for the simultaneous sensing of acceleration and capacitance [2]. Both designs support the detection of direct touch as well as proximity approach.

Plush Cube Design

The plush cube detects when the toy is touched and how tightly it is grasped. It consists of two concentric cubes, with the outer cube being much larger than the inner cube (see Figure 2). Each cube consists of furniture foam[1] covered with a single layer of conductive fabric. The inner cube and outer cube are wired to the *BlueSense* sensor using conductive threads and shielded coax cable. In order to avoid interference with the capacitance sensor and to prevent blocking the Bluetooth transmissions, the sensor package is placed inside a foam shelf that sits on top of the outer cube. The sensor battery is mounted outside of the foam shelf to facilitate easy charging as the sensor cannot be removed once it is connected. Both the foam shelf and concentric cubes are covered with a layer of flannel to conceal the shelf and prevent direct contact with the conductive fabric.

[1] From our previous caterpillar design, we know that the use of shape-retaining foam is essential to maintain a consistent baseline for capacitance.

Plastic Block Design

The plastic block interacts with Lego ™ Quatro blocks. It is constructed of ABS plastic and contains a sliding drawer to hold the sensor package. Two U-shaped copper sheets are inserted into opposing grooves inside the walls of the drawer. Holes in the drawer grooves expose two small sections of the copper plate and allow wire to be soldered directly from the copper plate to the *BlueSense* sensor (see Figure 3). The drawer can be reinserted without disturbing these connections.

Discussion

Both the cube and block toys were tested in play sessions. The plush cube was used in over 40 adult play sessions and seven child play sessions. The plastic block has been used in 25 adult play sessions and two child sessions. Both toys supported the detection of direct and indirect movements (see Figure 4). Despite its inability to detect grasp intensity, we favor the plastic block design. Connections inside the cube are brittle and required constant repair. The plastic block, thus far, is a more durable toy.

References

[1] Baranek, G. T., C. R. Barnett, E. M. Adams, N. A Wolcott, L. R Watson, and E. R Crais. "Object play in Infants with Autism: Methodological Issues in Retrospective Video Analysis." *American Journal of Occupational Therapy* 59, no. 1 (2005): 20-30.

[2] Presti, P. "BlueSense - A Wireless Interface Prototyping System." Master's Thesis, College of Computing, Georgia Tech, Atlanta, 2006.

[3] T. Westeyn, J. Kientz, T. Starner, and G. Abowd "Designing Toys with Automatic Play Characterization for Supporting the Assessment of a Child's Development" *Workshop on "Designing for Children with Special Needs" at* IDC 2008.

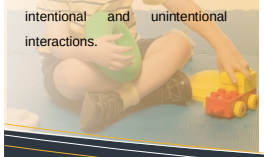
Capacitance Sensing in Smart Toys: Aiding the Detection of Play Behaviors

USING THE CHILDSPLAY SYSTEM

Georgia Tech College of Computing
School of Interactive Computing

MOTIVATION

- Certain aspects of object play have been linked to the potential early diagnosis of autism.
- We have investigated the use of sensors within toys to aid in the automatic analysis of play.
- Playtime can be very chaotic with toys unintentionally toppled, kicked, and knocked down.
- Often, acceleration data does not provide enough information to distinguish play activities from toys being accidentally bumped.
- Detecting the proximity of the child through capacitive sensing may help distinguish between the intentional and unintentional interactions.



OBJECTIVES

To explore the implementation and sensing capabilities of two augmented toys designed to distinguish between a child's direct manipulations and motions caused by kicks, accidental bumps, and other indirect interactions that can occur during play.

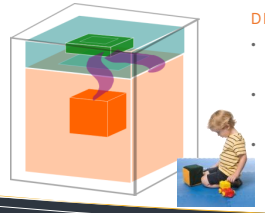
Supporting Rapid Analysis of Play Data

PLUSH CUBE TOY DESIGN

- The plush cube detects when the toy is touched and how tightly it is grasped.
- Two concentric foam cubes are each covered with a layer of conductive fabric.
- The sensor is encased in a foam shelf on top of the concentric cubes and is attached using conductive thread and a shielded coax cable.
- The toy is covered with flannel to prevent direct contact with the conductive fabric.

PLASTIC BLOCK TOY DESIGN

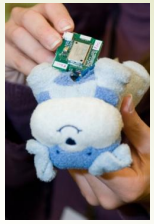
- The block interacts with Lego ™ Quatro blocks and is constructed of ABS plastic.
- It detects direct touches and approaches.
- The sensor is concealed in a drawer that has copper sheets inside opposing walls.
- The sensor is connected to the plate via two small holes in the drawer.
- The drawer can be reinserted without disturbing these connections.



DISCUSSION

- Both toys help support the detection of direct and indirect interactions and were each used in over 20 play sessions.
- Connections inside the cube are brittle and frequently broke during sessions when children sat or fell on the cube.
- Capacitive sensing plus acceleration helps discern accidental topples from the playfully knocking over of toys and offers a promising solution for recognizing object play.

THE CHILDSPLAY SYSTEM



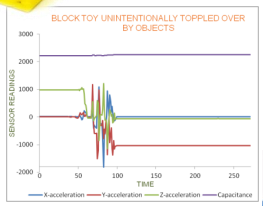
WIRELESS SENSORS INSIDE TOYS

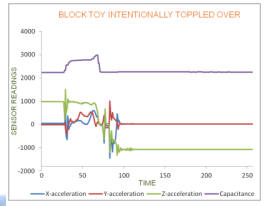
Toys transmit motion and touch information as children and adults play to a nearby computing platform

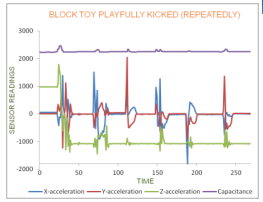
AUTOMATIC PLAY RECOGNITION

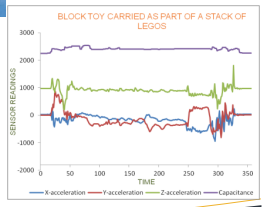
Various types of play, different levels of play sophistication, and the toys involved can be automatically identified and associated with video

PRELIMINARY RESULTS FROM PLAY TESTS









GvU GEORGIA TECH

Tracy L. Westeyn turtle@gatech.edu
Peter W. Presti peter.presti@imtc.gatech.edu
Jeremy M. Johnson Jeremy.johnson@imtc.gatech.edu
Scott Gilliland scott.gilliland@cc.gatech.edu
Jiasheng He jiazhe@cc.gatech.edu
David Quigley quigs@gatech.edu
Thad E. Starner thad@cc.gatech.edu

Interactive MEDIA Technology CENTER

This research has been supported by

AUTISM SPEAKS · Children's Healthcare of Atlanta · ARCS · NSF · Google · hsi Georgia Tech / Emory Initiative

APPENDIX G

CODING MANUAL FOR ADULT AND CHILD AUGMENTED–TOYS ONLY PLAY STUDY

actions	Description
sync	all toys in bag and shaken mutiple times (at start and end of sessions)
takeout	toys removed from bag (or pile of toys)
putaway	toy placed into the bag
explore	examines toy visually while manipulating it
fidget	actions inbetween scripted sequence
bang	single toy strikes surface 1 or more times
bump	toy moved unintentionally by another toy or body part
assemble	two toys fitted together. Lego-lego, dome-dome, red dome-ring
separate	assembled objects being pulled apart
relocate	toy at rest (on surface or held) is moved and brought back to rest
slide	objects move across surface while remaining in contact with surface
relate	two objects interacting with each other (ie during assembly)
reverb	result of previous interaction vibrating object when no longer touched
rock	dome tipped from one side to the other repeatedly.
roll	object tumbling across surface
shake	object rapidly moved up and down in the air
spin	object revolving about an axis perpendicular to surface
flies	object moves through air in a cyclic fashion (can be nested in a dome or ring)
toss	object thrown up in air and caught, or thrown back and forth between hands
runs	puppy used appropriately a moves from point A to point B
runs towards self	puppy runs towards the child playing with the toy
jumps	puppy moved over a stationary object, usually preceeded by running
pushes	**puppy** slides an object across the floor
rams	puppy runs towards an object and hits it with force (should include the entire movement of the other object)
wears	green ring is placed on the wrist and displayed
drinks	a single dome is used like a cup, picked up, sipped, and placed down
nests	toy is picked up and placed inside either dome or the ring. Puppy sitting on the ring considered a nest
hammers	the green ring repeatedly strikes the puppy (in a banging fashion)
spikes	puppy is droped or forcible thrown down to the ground head first
thrashes	the puppy is grabbed **BY THE HEAD** and hammered against other toys

Free Play Activities	
knockdown	bump causing drastic change in orientation
stack	one or more objects balanced ontop of eachother
drop	toy in hand falls to surface
falls	balanced toy falls off stack (or other toy) with no direct interaction
breaks apart	assembled ball comes apart
carries	toy held as child moves from point A to point B
crush	soft toys are forced inside other toys or sat upon
presents	toy is held in the direction of the researcher for veiwing
touches	hand rests on stationary toy anis not put in motion

	Exploration of Objects in Play						Relational Use of Objects in Play		Functional Use of Objects				
	Indiscriminate Actions			Manipulations of Single Objects			Takes Combinations Apart	Presentation and General Combinations	Semi Object Directed	Semi Self Directed	Playing with Puppy in Motion		
Toys	**explore**	**shake**	**bang**	**rock**	**spin**	**roll**	**take apart**	**assemble**			**puppy nested-ground**	**puppy ground motion**	**puppy nested-air**
Ring	Find the yellow dot on the ring	Does the ring have a rattle?	use the ring to hammer in an imaginary nail	*hold puppy by the head and thrash everything*	can you make the ring spin around like a record?	can you roll the ring	build the flying saucer; take apart ring and red dome	have trouble but put the red dome on the green ring	build the flying saucer and have it fly	put on the braclet	sit the puppy in the ring	have the puppy run towards the ring and jump over it	sit the puppy in the ring and fly him through the air
Dome (red)	Find the green dot on the red dome	Does the red dome have a rattle?	Use the red dome to hammer in an imaginary nail	can you make the red dome seesaw?	can you make the red dome spin?	can you roll the red dome like a tire?	build the ball; take apart the ball	have trouble but put the red dome on the blue dome	build the ball and roll it around	drink from the red cup	sit the puppy in the red dome, rock the dome	have the puppy run towards the red dome and jump over it	sit the puppy in the red dome and fly him through the air
Dome (blue)	Find the green dot on the blue dome	Does the blue dome have a rattle?	Use the blue dome to hammer in an imaginary nail	can you make the blue dome seesaw?	can you make the blue dome spin?	can you roll the blue dome?	build a tower with the domes; take apart stacked domes	have trouble but form a ball with the domes	build the ball and toss it inbetween your hands	drink from the blue cup	sit the puppy in the blue dome, rock the dome	have puppy run towards the blue dome and jump over it	sit the puppy in the blue dome and fly him through the air
Lego (yellow)	Find the blue dot on the lego	Does the lego have a rattle?	use the lego to hammer in an imaginary nail	*hold puppy by the head and thrash everything*	*spike the puppy face down like a football*	can you roll the lego	put the legos together; take apart legos	have trouble but put the legos together	0	put the legos together	have puppy push the lego like a bull dozer	have puppy run towards the yellow lego and jump over it	sit the lego in the blue dome and fly him through the air
Lego (grey)	Find the blue dot on the other lego	Does the other lego have a rattle?	use the other lego to hammer in an imaginary nail	*hold puppy by the head and thrash everything*	0	can you roll the other lego	put the legos together; take apart legos	have trouble but put the legos together	0	0	have puppy push the other lego like a bull dozer	have the puppy run towards the other lego and jump over it	sit the other lego in the blue dome and fly him through the air
Puppy	Find the red dot on the puppy	Does the puppy have a rattle?	use the puppy to hammer an imaginary nail	*hold puppy by the head and thrash everything*	0	can you roll the puppy	0	*hammer the puppy with the ring*	have puppy run around and bark	have the puppy run to you and bark	stack the legos on their side, have puppy run to it and knock them down	have the puppy run around and bark	have the puppy run to the lego and ram it
Block	Find the white dot on the block	Does the block have a rattle?	use the block to hammer an imaginary nail	*hold puppy by the head and thrash everything in site*	0	can you roll the dice	0	make a hamburger with the red dome, the green ring, and the blue dome	roll the dice	roll the dice	have puppy push the block like a bull dozer	have the puppy run towards the block and jump over it	have the puppy run to the other lego and ram it

APPENDIX H

CODING MANUAL FOR ADULT MIXED-TOY PLAY STUDY

Table 35: List of object play codes and definitions

Sheet1

actions	description	example
bang	single toy strikes surface	
bump	toy moved unintentionally by another toy or body part or shift in table	
drop	toy falls to surface or off surface	dropped into bag
grasp	hand around toy on table or holds toy steady off table	
join	two lego blocks fited together	
knockdown	bump causing drastic change in orientation	
manipulate	motion applied to toy while grasping	
move	toy at rest (on surface or held) is relocated and brought back to rest	
pickup	object lifted off a surface	
pour	contents of bag dumped onto table	
push	objects slid across surface	
putdown	object being held is placed on surface	
relate	two objects interacting with each other	
release	object is let go	
reverb	result of previous interaction shaking object when no longer touched	
roll	object tumbling across surface	
rub	object being stroked by an empty hand	petting the puppy
separate	lego blocks being pulled apart	
shake	object rapidly moved up and down in the air	
spin	object revolving across surface	lid or dome spun
spinning	wobble after and object is spung and no longer in contact with child	
stack	one object moved ontop of another object	lid placed on ring
takeout	objects moved from from a clutter of object or removed from a bag	remove from bag
unstack	object moved off of the object below it	

Page 1

The following is a list of the toys and objects to code names

Toys/Objects:
ring
puppy
lid
lego
dome
surface
stack

Label Scheme::

video is labeled using lower case letters with no spaces and follows the format:

action_object--notes

Action is one of 25 actions listed in the spreadsheet on the previous page (see Table 35). *Object* is one of the toys listed above. Actions and objects are separated by an underscore. An example label for a child banging the lid into the table would be:

bang_lid

If more than one object is interacting (ie: a toy in each hand) instead of listing the toy, a quantifying descriptor should be used. Valid quantifiers are:

Quantifiers:
two
many
all

An exmple for when to use quantifiers would be when two toys are being related or multiple toys are being pushed:

relate_two push_multiple

Notes are optional and are anything you wish to say about a specific label. They can indicate more descriptive information, they always follow a double hyphen. Single hyphens are used instead of spaces. Notes can be anything but will typically be one of the following:

Notes:
explore
relate

imaginary
object-object

For example when using a quantifier instead of a toy as the object portion of the label it is often useful to make a note about which objects are interacting. If a child were trying to fit the lego inside of the ring, the label could use the object-object note format and read:

relate_two--ring-lego

Note	**Description**
Explore	one toy being manipulated to learn about the object's properties
Relate	two objects interacting to discover properties of the objects
Imaginary	object(s) used in creative play (puppy dancing, eating from dome)

If a child made the puppy dance, it could be labeled as:

manipulate_puppy--imaginary-dance

If a child is searching for a feature on the dome, it could be labeled as:

manipulate_dome–explore

General Rules:

One label per activity: If two actions are occurring at the same time, label it as best you can with a single label

8x zoom start/stop: $8x$ zoom into the data in general and zoom from there to get start and stop as close as possible

SAVE EARLY, SAVE OFTEN!

REFERENCES

[1] Adamson, L. and Bakeman, R., "Viewing variations in language development: The communication play protocol," *Augmentative and Alternative Communication (Newsletter for ASHA Division 12)*, vol. 8, 1999.

[2] Adamson, L., Bakeman, R., and Deckner, D., "The development of symbol-infused joint engagement," *Child Development*, vol. 75, pp. 1171–1187, July/August 2004.

[3] Aeronautics, N. and Administration, S., "Nasa task load index (computer version): http://humansystems.arc.nasa.gov/groups/tlx/computer.php." World Wide Web electronic publication, Retrieved May 05, 2010.

[4] Agata Rozga, P. Personal Communication, 2009.

[5] Aylward, R., Lovell, S. D., and Paradiso, J. A., "A compact, wireless, wearable sensor network for interactive dance ensembles.," in *International Workshop on Wearable and Implantable Body Sensor Networks (BSN 2006), 3-5 April 2006, Cambridge, Massachusetts, USA*, pp. 65–70, IEEE Computer Society, 2006.

[6] Bao, L. and Intille, S. S., "Activity recognition from user-annotated acceleration data.," in *Pervasive Computing*, pp. 1–17, 2004.

[7] Baranek, G. T., Barnett, C., Adams, E., Wolcott, N., Watson, L., and Crais, E., "Object play in infants with autism: methodological issues in retrospective video analysis," *American Journal of Occupational Therapy*, vol. 59(1), pp. 20–30, 2005.

[8] Baranek, G. T., David, F. J., Poe, M. D., Stone, W. L., and Watson, L. R., "Sensory experiences questionnaire: discriminating sensory features in young children with autism, developmental delays, and typical development," *Journal of Child Psychology and Psychiatry*, vol. 47, no. 6, pp. 591–601, 2005.

[9] Blasco, P. A., "Pitfalls in developmental diagnosis," *Pediatric Clinics of North America*, vol. 38, pp. 1425–1438, 1991.

[10] Bricker, D. D., Squires, J., Potter, L. W., and Twombly, R. E., *Ages and Stages Questionnaires (ASQ): A Parent-Completed, Child-Monitoring System.* Paul H. Brookes Publishing CO, 1999. 6.

[11] Burges, C. J. C., "A tutorial on support vector machines for pattern recognition," *Data Min. Knowl. Discov.*, vol. 2, no. 2, pp. 121–167, 1998.

[12] Castelluccia, C. and Mutaf, P., "Shake them up!: a movement-based pairing protocol for cpu-constrained devices.," in *Proceedings of the 3rd International Conference on Mobile Systems, Applications, and Services (MobiSys 2005), June 6-8, 2005, Seattle, Washington, USA* (Shin, K. G., Kotz, D., and Noble, B. D., eds.), pp. 51–64, ACM, 2005.

[13] CDC, "Act early campaign website: http://www.cdc.gov/ncbddd/autism/actearly/." World Wide Web electronic publication, Retrieved April 16, 2008.

[14] Chang, C.-C. and Lin, C.-J., *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[15] Cohen, J., "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[16] DARPA, "ASSIST BAA #04-38, http://www.darpa.mil/ipto/ solicitations/open /04-38_PIP.htm," November 2006.

[17] Fell, H. J., Delta, H., Peterson, R., Ferrier, L. J., Mooraj, Z., and Valleau, M., "Using the baby-babble-blanket for infants with motor problems: an empirical study.," in *Proceedings of the ACM Conference on Assistive Technologies, ASSETS 1994, Marina Del Rey, California, USA, October 31 - November 3, 1994*, pp. 77–84, ACM, 1994.

[18] Fell, H. J. and Ferrier, L. J., "A baby babble-blanket.," in *INTERCHI Adjunct Proceedings* (Ashlund, S., Mullet, K., Henderson, A., Hollnagel, E., and White, T. N., eds.), pp. 17–18, ACM, 1993.

[19] First, L. and Palfrey, J., "The infant or young child with developmental delay," *The New England Journal of Medicine*, vol. 330, pp. 478–483, 1994.

[20] Fitzmaurice, G., *Graspable User Interfaces*. PhD thesis, University of Toronto, 1996.

[21] Gandy, M., Westeyn, T., Brashear, H., and Starner, T., "Wearable systems design issues for the elderly and disabled," in *Smart Technology for Aging, Disability, and Independence* (Helal, A., Mokhtari, M., and Abdulrazak, B., eds.), vol. 2, Wiley, In Press 2007.

[22] Ganesan, M., Russell, N. W., Rajan, R., Welch, N., Westeyn, T. L., and Abowd, G. D., "Grip sensing in smart toys: a formative design method for user categorization," in *CHI EA '10: Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*, (New York, NY, USA), pp. 3745–3750, ACM, 2010.

[23] Gorbet, M. G., Orth, M., and Ishii, H., "Triangles: Tangible interface for manipulation and exploration of digital information topography," in *Proceedings of CHI '98*, pp. 49–56, ACM, 1998.

[24] Groth-Marnat, G., *Handbook of Psychological Assessment, 3rd ed.* New York: John Wiley and Sons, 1997. 6.

[25] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H., "The weka data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, 2009.

[26] Hayes, G. R., Gardere, L. M., Abowd, G. D., and Truong, K. N., "Carelog: a selective archiving tool for behavior management in schools.," in *Proceedings of the 2008 Conference on Human Factors in Computing Systems, CHI 2008, 2008, Florence,*

Italy, April 5-10, 2008 (Czerwinski, M., Lund, A. M., and Tan, D. S., eds.), pp. 685–694, ACM, 2008.

[27] Hayes, G. R., Truong, K. N., Abowd, G. D., and Pering, T., "Experience buffers: a socially appropriate, selective archiving tool for evidence-based care," in *CHI '05: CHI '05 extended abstracts on Human factors in computing systems*, (New York, NY, USA), pp. 1435–1438, ACM Press, 2005.

[28] HouseN, "House nn website: http://architecture.mit.edu/housen/." World Wide Web electronic publication, Retrieved July 11, 2008.

[29] HTK, "HTK Speech Recognition Toolkit. Machine Intelligence Laboratory, Cambridge University. http://htk.eng.cam.ac.uk," 2007.

[30] Jurafsky, D. and Martin, J. H., *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Upper Saddle River, NJ, USA: Prentice Hall PTR, 2000.

[31] Kaelbling, L. P. and Saffiotti, A., eds., *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, UK, July 30-August 5, 2005*, Professional Book Center, 2005.

[32] Karam, M. and Schraefel, M. M. C., "Investigating user tolerance for errors in vision-enabled gesture-based interactions.," in *Proceedings of the working conference on Advanced visual interfaces, AVI 2006, Venezia, Italy, May 23-26, 2006* (Celentano, A., ed.), pp. 225–232, ACM Press, 2006.

[33] Kehoe, C., Cassell, J., Goldman, S., Dai, J., Gouldstone, I., MacLeod, S., O'Day, T., Pandolfo, A., Ryokai, K., and Wang, A., "Sam goes to school: story listening systems in the classroom," in *ICLS '04: Proceedings of the 6th international conference on Learning sciences*, pp. 613–613, International Society of the Learning Sciences, 2004.

[34] Kern, N., Antifakos, S., Schiele, B., and Schwaninger, A., "A model for human interruptability: Experimental evaluation and automatic estimation from wearable sensors.," in *8th International Symposium on Wearable Computers (ISWC 2004), 31 October - 3 November 2004, Arlington, VA, USA*, pp. 158–165, IEEE Computer Society, 2004.

[35] Kernberg, P. F., Chazan, S. E., and Normandin, L., "The children's play therapy instrument (cpti): Description, development, and reliability studies," *Journal of Psychotherapy Practice and Research*, vol. 7, pp. 196–207, July 1998.

[36] Kientz, J., Hayes, G., Westeyn, T., Starner, T., and Abowd, G., "Pervasive computing and autism: Assisting caregivers of children with special needs," *Special Issue on Pervasive Computing in Healthcare*, vol. Jan-Mar, 2007.

[37] Kientz, J. A., *Decision Support for Caregivers through Embedded Capture and Access.* PhD thesis, College of Computing, School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA, USA, 2008.

[38] Kientz, J. A., Arriaga, R. I., Chetty, M., Hayes, G. R., Richardson, J., Patel, S. N., and Abowd, G. D., "Grow and know: understanding record-keeping needs for tracking the development of young children," in *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, (New York, NY, USA), pp. 1351–1360, ACM Press, 2007.

[39] Kientz, J. A., Boring, S., Abowd, G. D., and Hayes, G. R., "Abaris: Evaluating automated capture applied to structured autism interventions.," in *Ubicomp* (Beigl, M., Intille, S. S., Rekimoto, J., and Tokuda, H., eds.), vol. 3660 of *Lecture Notes in Computer Science*, pp. 323–339, Springer, 2005.

[40] Kitamura, Y., Itoh, Y., and Kishino, F., "Real-time 3d interaction with activecube," in *CHI '01: CHI '01 extended abstracts on Human factors in computing systems*, (New York, NY, USA), pp. 355–356, ACM Press, 2001.

[41] Landis, J. R. and Koch, G. G., "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, pp. 159–174, 1977.

[42] Lena, "Lena website: http://www.lenababy.com/." World Wide Web electronic publication, Retrieved May 05, 2008.

[43] Lester, J., Hannaford, B., and Borriello, G., "'are you with me?' - using accelerometers to determine if two devices are carried by the same person," in *Proceedings of the Second International Conference on Pervasive Computing*, (Vienna, Austria), pp. 33–50, 2004.

[44] Lester, J., Choudhury, T., Kern, N., Borriello, G., and Hannaford, B., "A hybrid discriminative/generative approach for modeling human activities," in *Nineteenth International Joint Conference on Artificial Intelligence*, pp. 766–772, July 30 - August 5 2005.

[45] Lezak, M. D., *Neuropsychological Assessment*. New York: Oxford UP, 1983. 9.

[46] Lukowicz, P., Ward, J. A., Junker, H., Stäger, M., Tröster, G., Atrash, A., and Starner, T., "Recognizing workshop activity using body worn microphones and accelerometers.," in *Pervasive Computing, Second International Conference, PERVASIVE 2004, Vienna, Austria, April 21-23, 2004, Proceedings* (Ferscha, A. and Mattern, F., eds.), Lecture Notes in Computer Science, pp. 18–32, Springer, 2004.

[47] Lyons, K., Brashear, H., Westeyn, T., Kim, J. S., and Starner, T., "Gart: The gesture and activity recognition toolkit.," in *HCI (3)* (Jacko, J. A., ed.), vol. 4552 of *Lecture Notes in Computer Science*, pp. 718–727, Springer, 2007.

[48] Mayrhofer, R. and Gellersen, H., "Shake well before use: Authentication based on accelerometer data.," in *Pervasive Computing, 5th International Conference, PERVASIVE 2007, Toronto, Canada, May 13-16, 2007, Proceedings* (LaMarca, A., Langheinrich, M., and Truong, K. N., eds.), vol. 4480 of *Lecture Notes in Computer Science*, pp. 144–161, Springer, 2007.

[49] Minnen, D., Westeyn, T., Presti, P., Ashbrook, D., and Starner, T., "Recognizing soldier activities in the field," in *Proceedings of the Fourth International IEEE Workshop on Wearable and Implantable Body Sensor Networks (BSN 2007)*, pp. 236–241, March 26-28 2007.

[50] Minnen, D., Westeyn, T., Starner, T., Ward, J., and Lukowicz, P., "Performance metrics and evaluation issues for continuous activity recognition," in *Performance Metrics for Intelligent Systems*, 2006.

[51] Müller, H., Müller, W., Marchand-Maillet, S., and Pun, T., "Performance evaluation in content-based image retrieval: Overview and proposals," tech. rep., University of Geneve, Switzerland, 1999.

[52] of South Carolina, M. U., "Diagnostic tests glossary." Web Article, June 2006.

[53] Pallant, J. F., *SPSS survival manual : a step by step guide to data analysis using SPSS*. Allen and Unwin, Crows Nest, N.S.W. :, 3rd ed. ed., 2007.

[54] Patterson, D. J., Fox, D., Kautz, H. A., and Philipose, M., "Fine-grained activity recognition by aggregating abstract object usage," in *Ninth IEEE International Symposium on Wearable Computers (ISWC 2005), 18-21 October 2005, Osaka, Japan*, pp. 44–51, IEEE Computer Society, 2005.

[55] Platt, J., "Probabilistic outputs for support vector machines and comparisons to r egularized likelihood methods," in *Advances in Large Margin Classifiers*, MIT Press, 1999.

[56] Presti, P., "Bluesense - a wireless interface prototyping system," Master's thesis, College of Computing, Georgia Institute of Technology, Atlanta, GA, 2006.

[57] Rabiner, L., "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, pp. 257–286, Feb 1989.

[58] Rijsbergen, C. J. V., *Information Retrieval, 2nd Ed.* Butterworth Scientific Ltd., 1979.

[59] Rosa Arriaga, P. Personal Communication, 2007.

[60] Schapire, R. E., "A brief introduction to boosting," in *IJCAI*, pp. 1401–1406, 1999.

[61] Shannon, C. E., "Communication in the presence of noise," *Proceedings of the IRE*, vol. 37, no. 1, pp. 10–21, 1949.

[62] Sharlin, E., Itoh, Y., Watson, B., Kitamura, Y., Sutphen, S., and Liu, L., "Cognitive cubes: a tangible user interface for cognitive assessment," in *CHI '02: Proceedings of the SIGCHI conference on Human factors in computing systems*, (New York, NY, USA), pp. 347–354, ACM Press, 2002.

[63] Shevell, M., Ashwal, S., Donley, D., Flint, J., Gingold, M., Hirtz, D., Majnemer, A., Noetzel, M., and Sheth, R., "Practice parameter: Evaluation of the child with global developmental delay: Report of the quality standards subcommittee of the american academy of neurology and the practice committee of the child neurology society," *Neurology*, vol. 60, pp. 367–380, 2003.

[64] Stäger, M., Lukowicz, P., Perera, N., von Büren, T., Tröster, G., and Starner, T., "Soundbutton: Design of a low power wearable audio classification system.," in *7th International Symposium on Wearable Computers (ISWC 2003), 21-23 October 2003, White Plains, NY, USA*, pp. 12–17, IEEE Computer Society, 2003.

[65] Thelen, E., "Motor development as foundation and future of development psychology," *Journal of Behavioral Development*, vol. 24, pp. 385–397, 2000.

[66] Ullmer, B. and Ishii, H., "Emerging frameworks for tangible user interfaces.," *IBM Systems Journal*, vol. 39, no. 3-4, pp. 915–, 2000.

[67] Viola, P. and Jones, M., "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition*, 2001.

[68] Wang, P., Abowd, G. D., and Rehg, J. M., "Quasi-periodic event analysis for social game retrieval," in *Proceedings of IEEE International Conference on Computer Vision*, IEEE, 2009.

[69] Wang, P., Abowd, G. D., Rehg, J. M., and Arriaga, R. I., "Automatic retrieval of mother-infant social games from unstructured videos," in *Electronic Proceedings of the International Meeting for Autism Research (IMFAR)*, 2009.

[70] Wang, P., Westeyn, T., Abowd, G. D., and Rehg, J. M., "Automatic classification of parent–infant social games from videos," in *Electronic Proceedings of the International Meeting for Autism Research (IMFAR)*, 2010.

[71] Ward, J., Lukowicz, P., and Tröster, G., "Evaluating performance in continuous context recognition using event-driven error characterisation," in *Proceedings of LoCA 2006*, pp. 239–255, May 2006.

[72] Ward, J., Lukowicz, P., Tröster, G., and Starner, T., "Activity recognition of assembly tasks using body-worn microphones and accelerometers," *Pattern Analysis and Machine Intelligence (in press)*, 2006.

[73] Westeyn, T., Brashear, H., Atrash, A., and Starner, T., "Georgia Tech Gesture Toolkit: supporting experiments in gesture recognition.," in *Proceedings of the 5th International Conference on Multimodal Interfaces, (ICMI 2003)*, pp. 85–92, ACM, November 5-7 2003.

[74] Westeyn, T., Kientz, J., Starner, T., and Abowd, G., "Designing toys with automatic play characterization for supportin g the assessment of a child's development," in *Workshop on "Designing for Children with Special Needs" at the Se venth Conference on Interaction Design for Children (IDC 2008)*, June 11-13 2008.

[75] Westeyn, T., Presti, P., Arriaga, R., Starner, T., and Abowd, G., "An initial investigation using augmented toys and statistical mod els to automatically categorize object play behaviors," in *International Meeting of Autism Researchers (IMFAR 2009)*, May 07-09 2008.

[76] Westeyn, T., Presti, P., and Starner, T., "A nave technique for correcting time-series data for recognition applications," in *Proceedings of the Thirteenth IEEE International Symposium on Wea rable Computers (ISWC 2009)*, IEEE Computer Society, Sept 04-07 2009.

[77] Westeyn, T., Vadas, K., Bian, X., Starner, T., and Abowd, G. D., "Recognizing mimicked autistic self-stimulatory behaviors using hmms.," in *Ninth IEEE Int. Symp. on Wearable Computers*, pp. 164–169, October 18-21 2005.